基于认知的

汉语 HANYU

计算语言学研究

袁毓林 著



从认知的视角来研究计算语言学，特别是中文信息处理的问题。对有兴趣了解或从事计算语言学研究的人很有启迪意义。

从认知的角度研究了汉语的论元结构和描述框架，并进行了真实文本语义标注的实践。

结合作者自己的研究实践讨论说明了基于认知并面向计算的汉语语法研究的路线，展示了认知语言学和计算语言学相互结合的可能性。


ISBN 978-7-301-14052-9

9 787301 140529 >

定价：30.00元

基于认知的
汉语计算语言学研究

袁毓林 著



图书在版编目(CIP)数据

基于认知的汉语计算语言学研究/袁毓林著.—北京：北京大学出版社，2008.7

ISBN 978-7-301-14052-9

Ⅰ.基… Ⅱ.袁… Ⅲ.汉语-机器翻译-研究 Ⅳ.H085

中国版本图书馆 CIP 数据核字(2008)第 103402 号

书　　　名：基于认知的汉语计算语言学研究
著作责任者：袁毓林　著
责 任 编 辑：杜若明
标 准 书 号：ISBN 978-7-301-14052-9/H・2028
出 版 发 行：北京大学出版社
地　　　址：北京市海淀区成府路 205 号　100871
网　　　址：http://cbs.pku.edu.cn
电 子 信 箱：zpup@pup.pku.edu.cn
电　　　话：邮购部 62752015　发行部 62750672　编辑部 62752028
出版部 62754962
印　刷　者：北京大学印刷厂
经　销　者：新华书店
890 毫米×1240 毫米　A5　15.125 印张　335 千字
2008 年 7 月第 1 版　2008 年 7 月第 1 次印刷
定　　　价：30.00 元

陆　序

在学术研究领域，袁毓林可以说是一位勤奋的耕耘者。他的论文集《汉语语法研究的认知视野》（商务印书馆）于2004年出版，现在又推出了新的论文集《基于认知的汉语计算语言学研究》。我大略地翻阅了一下全书各篇的内容，论文集的书名"基于认知的汉语计算语言学研究"，点明了该书的基本内容——从认知的视角来研究计算语言学，特别是中文信息处理的问题。正文具体分四部分内容：

第一部分内容，作者取名为"计算理论和语言研究"，包括四篇文章：《计算语言学的理论方法和研究取向》、《基于统计的语言处理模型的有用性和局限性》、《认知科学和汉语计算语言学》和《面向当代科技的语言研究的理论和方法》。计算语言学的研究，大致可以分为两个层面，一个是理论模型的研究，一个是工程研究（或说具体的技术方法研究）。据我所知，袁毓林主要从事理论模型的研究，所以这一部分内容作者主要从宏观的角度介绍说明了计算语言学的理论方法和研究取向；评述了在自然语言处理中已运用过的基于规则和基于统计的两种处理模型，指出处理语言这种复杂的系统"必须走规则和统计相结合的道路"；从认知科学的视角作者把自己认为有价值的并且是可行的计算语言学研究模式介绍给读者，并结合作者自己的研究实践讨论说明了基于认知并面向计算的汉语语法研究的路线；展示了认知语言学和计算语言学相互结合的可能性。这部分内容对有兴趣了解或从事计算语言学研究的人来说，是值得一读的，是很有启迪意义的。

第二部分内容，作者取名为"论元结构和描述框架"，也包括四篇文章：《论元角色的层级关系和语义特征》、《一套汉语动词的论元角色的语法指标》、《汉语谓词的论元结构的描述框架》和《论元结构和句式结构互动的动因、机制和条件——表达精细化对动词配价和句式构造的影响》。袁毓林是我国最早研究配价问题的学者之一，特别是他第一个发表了有关汉语名词配价的研究成果，该成果被广为引用。

以乔姆斯基为代表的生成语法学派所提出的动词论元结构理论与法国依存语法学派特斯尼耶尔提出的动词配价结构理论有相同的一面，当然出发点不同，思考的角度不同，对语言事实解释的广度与深度也不同。十多年来袁毓林一直致力于动词论元结构的研究，在这方面他发表了一系列有分量的文章。我所主持的两个重大科研项目“面向中文信息处理的现代汉语动词论旨结构系统和汉语词语语义分类层级系统研究”（国务院 973 国家重点基础研究发展规划项目“图像、语音、自然语言理解与知识挖掘”子课题）和“汉语语义知识的形式化模型及语义分类系统研究”（教育部重点研究基地项目），袁毓林都参加了，其中的“汉语动词的题元系统及其语法指标”（包括“题元的层级体系”，“各别题元的定义、示例和句法语义特点”，“不同题元之间的配合关系”，以及“各别题元的语法指标”）就是由袁毓林执笔起草的。因此本书这一部分内容可以说是他对自己在配价问题和动词论元结构研究方面成果的汇集。在这部分内容中，他不仅建立并提出了汉语动词论元角色的层级体系，定义了各个语义角色，并细致描述了各个语义角色在述谓结构中所表现出来的动态性语义特征，同时通过十个各具特色、有代表性的实例（谓词“切、包$_1$、包$_2$、调查、帮忙$_1$、帮忙$_2$、飞$_1$、飞$_2$、吃、专政”）给出了谓词及其论元的句法配置方式，提出了汉语谓词论元结构的描写框架。更值得注意的是，他探讨了谓词论元结构和句式结构（constructions）互动的动因、机制和条件，对汉语谓词所谓“变价”和“论元增容”作了进一步的解释。

第三部分内容，作者取名为“信息抽取和语义标注”，包括五篇文章：《信息抽取的语义知识资源研究》、《用动词的论元结构跟事件模板相匹配——一种由动词驱动的信息抽取方法》、《用逻辑和篇章知识来约束模板匹配——逻辑结构和篇章结构知识在信息抽取中的运用》、《基于论元结构的语义标注的体系和规范》以及《新闻语体真实文本的语义标注的实践》。这部分内容作者主要提出并举例说明了要使计算机有效地自动从真实文本抽取信息，至少要有三种层面的语义知识：话语篇章知识、谓词论元结构知识和句子的逻辑结构知识；为对真实文本进行语义分析和标注，作者细致分析设计了篇章、谓词论元结构、句子逻辑结构这三种层面各自的语义关系，并为这三种层面各自的语义关系设计

并提出了一套可扩充的标记集;作者还以自己设计的这套标记对新闻报道中关于职务调动的真实文本进行了语义关系标注实践。作者标注得相当认真。通过这样的标注实践又有所发现——真实文本中代词或指示词的先行成分(一般称为先行语)常常是隐含的;段落之间的衔接,其形式手段相当缺乏。这就促使大家去进一步思考、探索怎么为计算机自动处理真实文本解决这方面的难题。

第四部分内容,作者取名为"专题研究和个案分析",也包括五篇文章:《容器隐喻和套件隐喻及相关的语法现象——词语同现限制的认知解释和计算分析》、《关于分词规范和规范词表的若干意见》、《中文信息处理中的语言难题问答》、《缓冲式移动通信及其发展方向——一个语言学家的设计思想》和《走向多层面互动的汉语研究》。这部分值得细细阅读的是《容器隐喻和套件隐喻及相关的语法现象——词语同现限制的认知解释和计算分析》和《走向多层面互动的汉语研究》这两篇文章。前一篇文章主要通过对"满"、"全",特别是"满+NP"、"全+NP"在意义、用法上的不平行性的解释,说明语言中的许多现象只有从认知的隐喻的视角来加以解释——用容器隐喻来解释"满"背后的概念结构以及由"满"构成的"满+NP"的使用特点,用套件隐喻来解释"全"背后的概念结构以及由"全"构成的"全+NP"的使用特点,这样才能说得清楚,说得圆满,说得充分,才能有解释力;通过对"满"和"全"又具有一定的可替换性的解释,说明隐喻分析有必要提升到更为抽象的意象图式水平,这样才更有解释力,才能最终解释说明既然"满"、"全"背后的概念结构是属于不同的隐喻范畴,为什么有时又具有可替换性,即才能说明为什么容器隐喻和套件隐喻在语言的实际使用中会出现二者中和化的现象;更积极的意义,还在于正如作者在文章中所指出的,有助于语言的认知解释有可能实现形式化和可计算,从而有可能实现认知和计算的统一("有可能"三个字不是作者说的,是我加的)。后一篇文章是作者为徐杰所编的《词汇语法语音的相互关联——第二届肯特岗国际汉语语言学圆桌会议(2002.11.26—30.)论文集》所写的代前言。文章扼要回顾了20世纪汉语研究的历史,对今后的汉语研究发表了很有见地的看法。作者强调指出,汉语研究必须树立"互动观念",走多层面互动研究之路,而这方面正

是目前汉语学界所缺乏的。文章特别谈到了一段时间来成为人们热门话题的所谓"语法研究三个平面"的问题,作者强调指出,"我们不仅应该分清语法的三个不同的平面,而且应该观察这三个不同的平面之间的互动关系",并应"引入语言类型学的视野","引进语法化这种动态性的概念,来审视语法、语义和语用这三个平面之间的互动关系","从而打破共时研究和历时研究之间的藩篱,把语言的共时研究和历时研究沟通起来",以"推动语言研究走向更为全面、综合和多层面互动的道路"。文章以学界已有的研究成果和作者本人的研究成果具体说明了语法和语音之间、词库结构和句法操作之间的互动关系,以及这种互动所应有的限度。这是很有见地的看法,应引起大家重视。

我虽然只粗粗阅读了一遍,觉得收获良多,推荐大家一读。借此机会我也想发表两点看法,同时也想提出一些意见。

第一点,当今语言研究的走向之一,确实如本书作者所说,要走多层面互动的研究之路。不过这只是"之一",还应有另一个"之一",那就是"特征研究",这也必须重视。从上个世纪七十年代以来,就语言研究说,一个重要的趋向是逐步重视特征的研究和描写。在语言的理论研究和应用研究上都是这样。

先说语言的理论研究,大家知道,在语言研究领域,最早讲特征的是音位学,接着是语义学;语法学里讲语义特征那是七十年代以后的事了。当时把"语义特征"这个概念术语借用到语法学中,为的是做两件事:一件事,用以解释造成同形多义句法格式的原因;另一件事,用以说明在某个句法格式中,为什么同是动词,或同是形容词,或同是名词,而有的能进入,有的不能进入。发展到乔姆斯基的生成语法理论,特征又赋予它新的含义。我们知道,乔姆斯基因为认为结构主义对语言的描写所概括的规则太复杂了,所以他要提出生成语法的观点,以简化语法规则。简约,一直是生成语法学的一个很重要的原则。从 1957 年的由核心句到非核心句的转换,到 1964 年的从深层结构到表层结构的转换,到上个世纪 80 年代初的 GB 理论——只剩下"α 移位"规则,其他都成了原则,再到最简方案及其近几年的论述——众多的原则和移位规则基本都不要了,D-结构,S-结构都没有

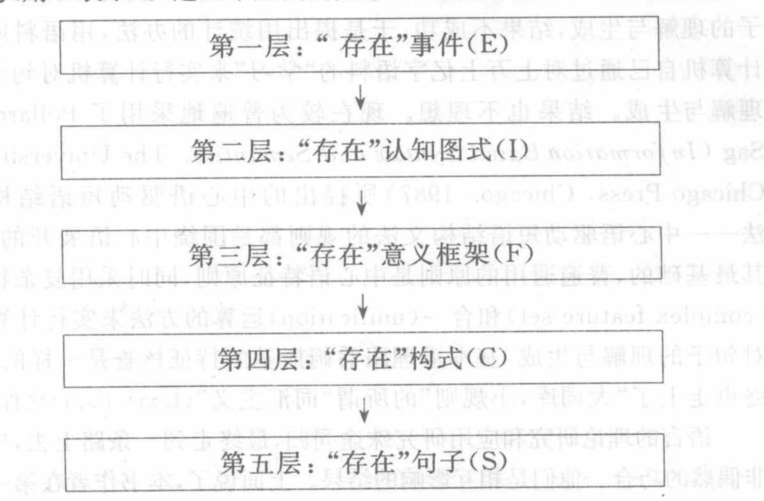
了，似只保留了"原则和参数"理论和"X-bar"结构模式，进一步强调经济原则；而提出了中心词(head)理论和特征核查(feature checking)理论，以及轻动词理论和VP空壳理论，注入了新的研究课题——接口(interfaces)的研究。基本的句法运作是从基础部分(即词库)取出带有各种各样的有关语义、句法特征的词项，进行来回合并(Merge)，如能通过特征核查(指中心语跟标示语，中心语跟补足语，在特征上吻合)，由此生成的词项组合再去跟音韵接口，跟逻辑语义接口，从而最终生成我们所听到或看到的句子。总之，词语的特征的分析和描写放到了非常重要的位置，走上了"大词库，小规则"之路。这里要附带说明的，最近乔姆斯基在 Linguistic Inquiry 杂志 2005 年第1期上发表的文章(*Three Factors in Language Design.*)中似乎提到要取消"特征核查"，但他同时在文章中认为，从词库选出词语项，构成词语序列，形成语段后，要通过所谓"探针(probe)"与目标进行相互核查，如果没有发现不可诠释的特征，就转移给语音和语义两个界面接口，由此获得语音和语义相结合的语言形式。这实质上还是需要进行特征核查这一步骤。而所谓"要取消特征核查"，我体会是指在操作手续上要进一步简化。

现在再看自然语言处理与理解这方面的语言应用研究。大家都知道，自然语言处理与理解最早使用规则的方法来实行计算机对句子的理解与生成，结果不成功；于是提出用统计的办法，用语料库让计算机自己通过对上万上亿字语料的"学习"来实行计算机对句子的理解与生成。结果也不理想。现在较为普遍地采用了 Pollard & Sag (*Information Based Syntax and Semantics*. The University of Chicago Press, Chicago. 1987)所提出的中心语驱动短语结构文法——中心语驱动短语结构文法的规则都是围绕中心语展开的，而其最基础的、普遍通用的原则是中心语特征原则，同时采用复杂特征(complex feature set)和合一(unification)运算的方法来实行计算机对句子的理解与生成，基本道理跟乔姆斯基的特征核查是一样的，最终也走上了"大词库，小规则"的所谓"词汇主义"(lexicalism)之路。

语言的理论研究和应用研究殊途同归，最终走到一条路上去，这绝非偶然的巧合。他们是相互影响的结果。上面说了，本书作者在第一部

分内容里，主要从宏观的角度介绍说明了计算语言学的理论方法和研究取向，但作者未注意到"重视特征研究"这一取向，这可能跟作者对于国外近十多年来有关计算语言学方面的文献资料还了解得不全面有关。

第二点，上面说了，作者在第二部分内容里，探讨了谓词论元结构和句式结构(constructions，有人说成"构式")互动的动因、机制和条件，对汉语谓词所谓"变价"和"论元增容"作了进一步的解释。论述很有新意。但我觉得如果作者能进一步深入思考这样一个问题就好了："人对客观事物的感知所得最后是怎样用言辞表达出来的?"最近看到王黎在《关于构式和词语的多功能性》(《外国语》2005 年第 4 期)一文中明确提出了这个问题。王黎认为，从人对客观事物的感知所得到最后用言辞表达出来，中间一共可分为五个层面：第一层：是客观世界中所存在的诸多基本的、典型的事件(包括景象等)，诸如"存在事件"、"分配事件"、"事物特征"等(用 E 表示)；第二层：这个事件，如存在事件，被人观察到以后，就相应地在人的认知域里，形成了存在意象(image，用 I 表示)；第三层：这存在意象又激活了人脑里的深层存在意义框架(用 F 表示)；第四层：当这个深层存在意义框架被位于表层的语言表现出来时，就有了存在构式(用 C 表示)；第五层：那存在构式里填上一定的具体的词项，就形成我们在实际语言交际中所听到看到的存在句(用 S 表示)。这五个层面的关系，王黎图示如下：



这当然是一种假设，不能看作是结论，但可引起人们去进一步思考。同时，可以用来更好的解释说明句式的配价问题，也可以对汉语谓词所谓“变价”和“论元增容”作出更好的解释。

袁毓林论文集编就后，要我写序，这已是数个月之前的事了。写序，还是应尽可能做到有的放矢，实事求是。所以我在动笔写之前，一定要先看书稿，了解全书内容。我又比较忙，这样就拖到现在才将序文草就。所言不一定到位，请作者和广大读者批评指正。是为序。

陆俭明

2007-10-08 于北京蓝旗营寓所

冯 序

读了袁毓林教授新著的文集《基于认知的汉语计算语言学研究》,使我联想到美国著名人工智能专家 T. Winograd 在 1983 年写的专著《作为认知过程的语言》(*Language as a Cognitive Process*)。这两本书都试图从认知的角度来研究计算语言学的问题。可惜 Winograd 的专著只写了“句法”(Syntax)部分,没有再继续往下写。几年以前,我在国外曾经遇见 Winograd,问他为什么不继续写“语义学”(Semantics)部分,他回答说,语义学太复杂,不打算继续写下去了。这样,《作为认知过程的语言》这本专著可以说只是写了一半,就半途而废了。从 Winograd 的学识和才气来说,他是完全可以继续写下去的;可是他没有继续写,我感到非常之可惜。毓林的这本文集,着重从认知的角度探讨论元结构和语义标注,基本上都是语义的问题,恰好弥补了 Winograd 专著的不足,令我感到兴奋。

T. Winograd 在他的专著中说,为了从认知的角度来研究语言,应该解决如下两个问题:

第一,一个人要说话和理解语言,必须具有哪些知识?

第二,为了在语言交际中使用这些知识,人们的心智(mind)是怎样组织这些知识的?

根据研究计算语言学多年的实践经验,一个人在说话和理解语言时,不仅需要关于语言的知识,而且还需要各种非语言的知识,例如关于外在世界的知识、日常生活中的常识等,这已经是不容争论的问题。事实上,计算语言学研究者也在努力把这些知识形式化,以便计算机处理。但是,要了解人们的心智究竟怎样组织这些知识,却是一个十分困难的问题。认知语言学试图解决这样的问题。

认知语言学是 20 世纪 80 年代才出现的语言学科,如果把 1989 年在德国 Duisburg 召开的国际第一届认知语言学会议作为认知语言学诞生的标志,那么,这门学科至今才有短短 19 年的历史,可以说

是非常年轻的学科。其实,在认知语言学产生之前,很早就有人提出了通过语言来揭示人类心智的问题,已经涉及到认知语言学的问题。1933年,英国数学家A. M. Turing就预见到未来的计算机将会对自然语言研究提出新的问题。他在《机器能思维吗》一文中指出:"我们可以期待,总有一天机器会同人在一切的智能领域里竞争起来。但是,以哪一点作为竞争的出发点呢?这是一个很难决定的问题。许多人以为可以把下棋之类的极为抽象的活动作为最好的出发点,不过,我更倾向于支持另一种主张,这种主张认为,最好的出发点是制造出一种具有智能的、可用钱买到的机器,然后,教这种机器理解英语并且说英语。这个过程可以仿效小孩子说话的那种办法来进行。"Turing提出,检验计算机智能高低的最好办法是让计算机来讲英语和理解英语,他天才地预见到计算机和自然语言将会结下不解之缘。我认为,Turing这种预见的实质,就是提出了"语言是认知的窗口"的这个重要命题。这个命题是认知语言学的基础。所以,从认知的角度来研究计算语言学,进行"基于认知的汉语计算语言学研究",是非常必要的。

毓林在这本文集中,从认知的角度研究了汉语的论元结构和描述框架,并进行了真实文本语义标注的实践,使我们对于汉语的论元结构有了更加深刻的认识。

在20世纪70年代末和80年代初,我在法国格勒诺布尔理科医科大学研制汉-法/英/日/俄/德多语言机器翻译系统FAJRA时,就根据Tesnière的依存语法(grammaire de dépendence),对汉语动词、形容词和部分名词的论元结构进行了初步的探索,当时我提出的论元有:施事、受事、与事、关涉、时刻、时段、时间起点、时间终点、空间点、空间段、空间起点、空间终点、初态、末态、原因、结果、工具、方式、目的、条件、作用、内容、范围、论题、修饰、比较、伴随、判断、陈述、附加等,共30个,其中,施事、受事、与事3个论元是"行动元"(actants),其他27个论元是"状态元"(circonstants)。我根据机器词典中存储的单词的语法和语义的静态信息以及在句法分析中运算得出的句法功能的动态信息,使用计算机求解了这些论元信息,把汉语自动地翻译成5种外语,顺利了完成了多语言机器翻译实验。可是,我在

20多年前对于汉语论元结构的研究，是从依存语法和工程应用的角度出发的，根本没有考虑到这些论元的认知基础。

现在，毓林从认知的角度，根据计算机处理汉语的实际需要，详细地研究了汉语动词论元结构的论元属性、论旨属性、语法特征、语义特征、配位方式，把汉语动词的论元分为施事、感事、致事、主事、受事、与事、结果、对象、系事、工具、材料、方式、场所、源点、终点、范围、命题，共17个。并且使用自立性、使动性、感知性、述谓性、变化性、受动性、渐成性、关涉性、类属性等动态语义特点以及句法特点，来区分这些论元，从而明确地界定了这些论元。毓林的研究，在更深的层次上揭示了汉语论元结构的特性和判断方法，在逻辑上很有魅力，使我们得到一种逻辑上的美感。但是，他提出的这17个论元中，没有表示时间、原因、目的、论题的论元，而这些论元，在真实的文本中是经常出现的；而且毓林提出的命题这个论元，实际上就是句子，显然是不必要的。

也许毓林察觉到了他的这个论元系统的不足，后来他在语料库语义标注的实践中，把他的这17个论元进一步做了扩充。增加了经事、原因、目的、时间、路径、话题、说明等论元，删除了原来的命题论元，共23个，形成了他的"论旨角色标记集"。这个标记集基本上覆盖了我原来的30个论元的标记集，而且更加精炼，每一个论元的区别特征也更加清楚了，我赞同并且非常欣赏毓林的这个标记集。

毓林把他的研究成果应用于新闻语体真实文本的语义标注和信息自动抽取，效果良好，证明了论元结构知识的广泛适用性。他的成功说明了认知语言学对于计算语言学的理论和实践确实是很有吸引力的。计算语言学应该吸取认知语言学的成果，从而促进自身的发展。

认知科学的基础是"物理符号系统假设"。这种假设认为，智能的基础是符号操作，一切认知系统本质上都是符号加工系统，而符号操作就是计算，认知就是计算。

早在80年代初期，著名语言学家J. A. Fodor在《表达》(*Representations*)一书(MIT Press, 1980)中就说过："只要我们认为心理过程是计算过程(因此是由表达式定义的形式操作)，那么，除了将

心智看作别的之外，还自然会把它看作一种计算机。也就是说，我们会认为，假设的计算过程包含哪些符号操作，心智也就进行哪些符号操作。因此，我们可以大致上认为，心理操作跟图灵机的操作十分类似。"Fodor 在这里所说的"符号操作"，实际上也就是"规则"，所以，这种说法代表了计算语言学中的基于规则的理性主义观点。这种理性主义的观点，完全被后来兴起的认知语言学继承并进一步发展了。

而在认知语言学产生之前，在计算语言学中的这种基于符号操作规则的理性主义的观点早就受到了学者们的批评。1980 年，J. R. Searle 在他的论文《心智、大脑和程序》(*Minds, Brains and Programmes*)(1980，载《行为科学与脑科学》[*Behavioral and Brain Sciences*], Vol. 3)中，提出了所谓"中文屋子"的质疑。他提出，假设有一个懂得英文但是不懂中文的人被关在一个屋子中，在他面前是一组用英文写的指令，说明英文符号和中文符号之间的对应和操作关系的种种规则。这个人要回答用中文书写的几个问题，为此，他首先要根据指令规则来操作问题中出现的中文符号，理解问题的含义，然后再使用指令规则把他的答案用中文一个一个地写出来。这显然是非常困难的而且几乎是不能实现的事情。Searle 的批评是非常尖锐的，这样的批评使计算语言学中基于符号操作规则的理性主义的观点受到了普遍的怀疑。

这种理性主义方法的另一个弱点是在实践方面的。计算语言学中的理性主义者往往把自己的目的局限于某个十分狭窄的专业领域之中，他们采用的主流技术是基于规则的句法-语义分析技术，尽管这些应用系统在某些受限的"子语言"中也曾经获得一定程度的成功，但是，要想进一步扩大这些系统的覆盖面，用它们来处理大规模的真实文本，仍然有很大的困难。因为从自然语言系统所需要装备的语言知识来看，其数量之浩大和颗粒度之精细，都是以往的任何系统所远远不及的。而且，随着系统拥有的知识在数量上和程度上发生的巨大变化，系统在如何获取、表示和管理知识等基本问题上，不得不另辟蹊径。这样，基于统计的经验主义方法就越来越受到计算语言学研究者的欢迎。

毓林的这本文集，尽管其主要内容是讲基于认知的汉语计算语

言学研究，但是，他也注意到了计算语言学中基于统计的经验主义方法，他直率地指出了基于统计的语言处理模型的“有用性”和“局限性”，并且认为，“语言信息处理面临的对象既然有如此顽劣的既抗拒规则模型、又抗拒统计模型的属性，那么一种可能的技术途径只能是把规则的方法和统计的方法结合起来”。很多认知语言学家都推崇认知理论而排斥统计方法，而毓林独具慧眼，他重视认知而不排斥统计，主张规则方法和统计方法的结合，这是难能可贵的。

毓林在他的文集中，非常推崇“计算语言学是用计算机和为计算机研究语言的学科”这个关于计算语言学的定义。并且说，这个定义是国际计算语言学界对计算语言学的定义逐步形成的“共识”。这种说法未免有些偏颇。

我认为，科学的定义应该揭示计算语言学这个学科的本质属性，而毓林所推崇的这个定义带有明显的实用色彩，没有反映出计算语言学与计算机科学在理论上的联系，因而也就难以反映这个学科的本质属性。如果一个人在研究语言时，仅仅使用计算机来统计某些语言单位的出现次数，显然还谈不上他是在研究计算语言学，尽管他用计算机研究了语言；同样地，如果一个人仅仅为了在计算机上输入汉字而研究汉字编码，显然也谈不上他是在研究计算语言学，尽管他是在为计算机研究语言。计算语言学是一个独立的学科，它不仅有着严格而系统的理论，而且还有着完善而成熟的方法，计算语言学的这些理论和方法，正如物理学、数学和化学的理论和方法一样，绝不是不学而能的，而是要经过刻苦的学习和反复的实践才能掌握的。如果一个语言学家只是使用计算机来研究语言而不懂计算语言学的基本理论和方法，他只是一个使用计算机的语言学家，还谈不上是一个计算语言学家；如果一个计算机专家为了在计算机上输入汉字来研究汉字编码而不懂得计算语言学的基本理论和方法，他也只是一个为计算机而研究语言的计算机专家，还谈不上是一个计算语言学家。

毓林说他推崇的这个定义已经逐渐成为国际计算语言学界的“共识”，可能与事实不符。我查阅了很多英文文献，并没有发现这个定义，我还查阅了法文、德文、俄文、日文的文献，也没有发现这个定

义。可见,这个定义远远还没有成为国际计算语言学的普遍共识。

如果我们把1954年第一次机器翻译实验的成功算做计算语言学的开始,那么,计算语言学这个学科已经有50多年的历史了,在计算语言学创始前后那个充满了理性的年代,计算机科学的先行者Turing和Shannon就非常重视计算机科学的理论和自然语言的联系。Turing提出了著名的Turing实验,认为检验计算机智能高低的最好办法是让计算机来讲英语和理解英语。Shannon在他的《通信的数学理论》(*Mathematical Theory of Communication*)中,用马尔可夫过程的理论来分析英语,建立了信息论的基础。他们独树一帜的研究都与自然语言有着千丝万缕的联系,他们的远见卓识都为计算语言学播下了科学的种子。50多年来,他们播下的种子早已破土而出,由纤细柔弱的嫩芽长成了枝叶茂密的大树,成为了一门独立的学科。所以,在给计算语言学这个学科下定义时,我们切不可忽视它与计算机科学在理论上的深刻联系,只有这样,才有可能揭示出这个学科的本质属性。

《计算机进展》(*Advanced in Computer*)是国际计算机科学的权威出版物,这个出版物登载的文章,都是引导计算机科学学术潮流的高质量论文,从中我们可以窥见国际计算机科学的发展方向。

美国计算机科学家Bill Manaris在1999年出版的《计算机进展》第47卷的《从人-机交互的角度看自然语言处理》一文中曾经给"自然语言处理"提出了如下的定义:

"自然语言处理可以定义为研究在人与人交际中以及在人与计算机交际中的语言问题的一门学科。自然语言处理要研制表示语言能力和语言应用的模型,建立计算框架来实现这样的语言模型,提出相应的方法来不断地完善这样的语言模型,根据这样的语言模型设计各种实用系统,并探讨这些实用系统的评测技术。"这个定义的英文如下:"NLP could be defined as the discipline that studies the linguistic aspects of human-human and human-machine communication, develops models of linguistic competence and performance, employs computational frameworks to implement process incorporating such models, identifies methodologies for iterative refine-

ment of such processes/models, and investigates techniques for evaluating the result systems."(Bill Manaris:〈Natural language processing: A human-computer interaction perspective〉, *Advances in Computers*, Volume 47,1999)

Bill Manaris 关于自然语言处理的这个定义,比较全面地表达了计算机对自然语言的研究和处理的主要内容,说明了自然语言处理不仅要研究表示语言能力(linguistic competence)的模型,而且还要研究表示语言应用(linguistic performance)的模型,涉及到了自然语言处理在理论上的本质问题,因此,这个定义在《计算机进展》上发表以后,逐渐得到国际自然语言处理界的共识。这个定义是针对"自然语言处理"而提出的,而"自然语言处理"与"计算语言学"是如此之接近,在这里,我愿意推荐这个定义给毓林,作为他给计算语言学这个学科下定义的参考。

计算语言学的研究范围涉及到众多的部门,如语音的自动识别与合成、机器翻译、自然语言理解、人机对话、信息检索、文本分类、自动文摘、机器词典、语料加工、算法研究、语言形式模型研究,等等。我们认为,这些部门可以归纳为如下四个大的方向:

■ 语言工程方向:把自然语言处理作为面向实践的、工程化的语言软件开发来研究。这一方向的研究一般称为"人类语言技术(Human Language Technique, 简称 HLT)",或者称为"语言工程"(Language Engineering)。

■ 数据处理方向:把自然语言处理作为开发语言研究相关程序以及语言数据处理的学科来研究。这一方向的研究早期的研究有术语数据库的建设、各种机器可读的电子词典的开发,近年来随着大规模语料库的出现,这个方向的研究显得更加重要。

■ 人工智能和认知科学方向:把自然语言处理作为在计算机上实现自然语言能力的学科来研究,探索自然语言理解的智能机制和认知机制。这一方向的研究与人工智能以及认知科学关系密切。

■ 语言学方向:把自然语言处理作为语言学的分支来研究,它

只研究语言及语言处理与计算相关的方面，而不管其在计算机上的具体实现。这个研究方向的最重要的研究领域是语法形式化理论和自然语言处理的数学理论。

我国的计算语言学研究在语言工程方向和数据处理方向已经投入了很多的资金和人力，大多数的计算语言学工作者都在探索这两个方向的问题，硕果累累。但是，对于人工智能和认知科学方向以及语言学方向，投入就比较少，研究的人也不多，显得比较薄弱。毓林的这本文集就是专门探讨这两个方向的各种理论和实践问题的，而且已经取得了煌煌的成绩，令我感到兴奋。我希望有更多的学者能够重视这两个方向的研究，弥补我国计算语言学研究的这些薄弱环节。

计算语言学是语言学、计算机科学和数学的交叉学科。每一个从事计算语言学研究的人，都面临着知识更新的问题。毓林是中文系出身的一个文科学者，为了研究计算语言学，他进行了知识更新的再学习，从他的文集中可以看出，他不但对于计算机科学和数学不是似懂非懂的外行，而且他还熟悉计算语言学本身独有的基本理论和方法，是计算语言学的精研通达的内行，他使用这些理论和方法，把计算机科学和数学的知识与语言研究有机地、巧妙地融为一体。毓林的研究，把文科与理科结合起来，把汉语与外语结合起来，把理论和实践结合起来，我相信，今后毓林在计算语言学的研究中，一定会做出更加出色的成绩。

冯志伟

于北京后拐棒胡同寓所

2007年11月10日

目 录

一、计算理论和语言研究

二、论元结构和描述框架

三、信息抽取和语义标注

四、专题研究和个案分析

五、附 录

一、计算理论和语言研究

计算语言学的理论方法和研究取向

本文从研究取向的角度，对目前计算语言学的几种理论方法以及相应的语言处理技术进行比较研究。着重讨论工程主义、工具主义、认知主义、实证主义和逻辑主义五种研究取向，比较几种关于人类知识和语言理解过程及相应的计算机模拟策略的理论，分析其在具体的语言处理技术（包括语法形式体系、语义表示体系、分析算法以至程序实现）上的差异。希望对不同的理论方法和处理技术的效能和局限有一个比较清楚的认识，从而为汉语计算语言学的研究提供借鉴。

0　计算机：语言研究的奴仆还是上帝

在社会语言学、文化语言学、心理语言学、神经语言学、认知语言学、数理语言学和计算语言学等当代带分号的语言学（hyphenated linguistics）中，计算语言学是一门跟当代科学技术关系最密切的学科，同时也是一门定义最为纷歧的学科。只要打开有关的文献，你就能找到关于计算语言学的各种差别极大的定义。事实上，这些不同的定义背后反映了不同的研究者的不同的研究取向。其中，最核心的一点是：怎样看待计算机和语言研究的关系，是把计算机作为语言研究的工具、还是作为语言研究的目标和服务对象。形象地说，把计算机当作为语言研究服务的奴仆、还是当作语言研究要为之服务的上帝。

下面，我们通过五种关于计算语言学的定义，来讨论工程主义、工具主义、认知主义、实证主义和逻辑主义五种不同的研究取向，比较不同的研究者为了实现这些不同的目标而采用的迥然不同的理论和方法（包括对人类知识、语言习得和语言理解过程的看法、以及相应的在计算机上模拟的策略），分析其在具体的语言处理技术（包括语法形式体系、语义表示体系、分析算法以至程序实

现)上的差异。希望对计算语言学中不同的理论方法和处理技术的效能和局限有一个比较清楚的认识,从而为汉语计算语言学的研究提供借鉴。

1 工程主义取向:着眼于计算机系统的建立

在计算语言学的诸多定义中,最多的是着眼于建立一种可运转的计算机系统。例如:

(1) Computational linguistics is the study of computer systems for understanding and generating natural language.

——Grishman (1986), p. 4

(计算语言学是对能理解和生成自然语言的计算机系统的研究)

(2) 计算语言学是采用计算机技术来研究和处理自然语言的一门新兴学科。

——冯志伟(1992),第 84 页

持这种观点的学者自然会把计算语言学的研究重点放在这种能理解和生成自然语言的计算机系统的结构及相应的各种算法的设计上。因为,从理论上说,要想让计算机去解决某种问题,必须满足下列三个基本的前提条件:①

第一,必须把待解的问题形式化。由于计算机只能对有限符号集上的有限长度的符号序列进行决定型的形式变换(这就是计算),因而首先要建立一个形式系统(formalism,一译形式体系):规定所用的各种符号(词汇),规定把符号连接成合法序列(即合式公式)的规则(句法),规定合法的符号串如何表示特定问题领域中的意义(语义,或解释);然后,建立一些推理规则,说明对这些符号和合法符号串可以进行一些什么样的处理(演算)。于是,问题便可以用符号表达出来,问题的解也表现为对符号序列的条件。这样,计算机解决问题的过程就是从表示问题的符号序列出发,按规则进行加工,一直到

① 详见马希文(1986),第 225—228 页。

得出符合要求的符号序列(即解)为止。这一整套的办法叫形式化(又叫数学方法),其要义是:把特定领域的问题转变为符号,从而把对问题的求解转变为对符号串的变换处理。

第二,这种问题必须是可计算的(computable),即一定要有解题的算法(algorithm),使得计算机能按照算法所指引的解题步骤,通过有限步的运算而得出结果。

第三,这种问题必须有一个合理的复杂度,也就是要避免指数爆炸(exponential explosion)。也就是说,问题的复杂性必须限制在目前的数字计算机的存储空间和运算时间所能容忍的范围之内。

所以,从研究程序上讲,这种类型的计算语言学研究一般分为如下三个阶段:[①]

> 第一步,数学建模。把需要研究的问题在语言学上加以形式化(linguistic formalism),使之能以一定的数学形式、严密而规整地表示出来。也就是说,为有关的语言问题建立数学模型。包括选择恰当的形式语法(formal grammar)使得句子的结构能够用某种数学形式明确而清晰地表示出来,研究在这种形式语法之下如何分析句子构造的方法和步骤;选择恰当的表示体系使得句子的意义能够用某种数学形式明确而清晰地表示出来,研究在这种形式体系之下如何分析和表示句子的语义结构。
>
> 第二步,算法设计。把这种严密而规整的数学形式表示为算法(algorithm),使之在计算上形式化(computational formalism)。这就必须研究句子分析的严格的手续(procedures),并抽象成机械的、明确的、一步步逼近分析结果的步骤。
>
> 第三步,程序实现。根据算法用某种程序语言编写计算机程序,使之在计算机上加以实现(computer implementation)。

比如,假定有下面这部小型的用产生式(production)表示的语境自由的短语结构语法:

① 参考冯志伟(1992),第84页;钱锋(1990),第26—27页。

S→NP+VP ……R1

NP→N ……R2

NP→PRO ……R3

VP→Vi ……R4

VP→Vt+NP ……R5

那么,句子 I like cheese.(我喜欢奶酪)的最左推导是:

S→NP+VP

→PRO+VP

→PRO+Vt+NP

→PRO+Vt+N

为了让计算机能根据上面给出的语法规则自动地分析这个句子,必须设计相应的算法:或者是自顶向下的回溯算法,或者是自底向上的并行算法。自顶向下的回溯算法每次只尝试一种推导,当一种推导失败时便返回、重新尝试另一种推导;就这样逐个地枚举语法所允许的各种推导,直至找到一个能生成输入句子的推导。根据这种算法(具体的细节从略),对于上文那部只有五条规则的语法,句子 I like cheese. 的推导过程将表现如下:

i. S

ii. S→NP+VP

iii. S→NP+VP→N+VP

iv. S→NP+VP→PRO+VP

v. S→NP+VP→PRO+VP→PRO+Vi

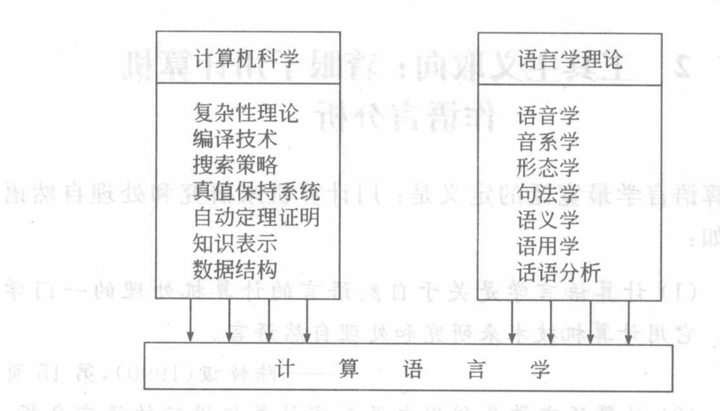
vi. S→NP+VP→PRO+VP→PRO+Vt+NP

vii. S→NP+VP→PRO+VP→PRO+Vt+N

i. S 是初始符,即树顶节点;ii. 根据 R1,展开初始符;iii. 根据 R2 展开最左的非终结符,但是范畴 N 跟词项 I 不匹配,需要回溯;iv. 根据 R3 展开最左的非终结符,范畴 PRO 跟词项 I 匹配成功;v. 根据 R4 展开左端第二个非终结符,但是范畴 Vi 跟词项 like 不匹配,需要回溯;vi. 根据 R5 展开左端第二个非终结符,范畴 Vt 跟词项 like 匹配

成功；vii. 根据 R2 展开最后一个非终结符，范畴 N 跟词项 cheese 匹配成功；至此，推导结束。[①]

一般地说，计算语言学的研究既必须涉及计算机科学中的复杂性理论（complexity theory，用以判别所研究的问题是否具有可计算性）、编译技术（compiler technology）、搜索策略（search strategies）、真值保持系统（truth-maintenance systems）、自动定理证明（automatic theorem proving）、知识表示（knowledge representation）和数据结构（data structure）等方面，同时也必须涉及语言学中的语音学（phonetics）、音系学（phonology）、形态学（morphology，或词法学）、句法学（syntax）、语义学（semantics）、语用学（pragmatics）、话语分析（discourse analysis）等方面。可以图示于下：[②]



如果说科学是理论和知识体系、技术是方法和操作技巧、工程是实践和具体施行的话，那么计算语言学就是一种工程。为了建造一个顺畅（fluent）、健壮（robust）的自然语言处理系统，必须整合许多不同类型的知识；诸如句法知识、语义知识、话语领域知识等，并且要有效地用到自然语言处理系统中。正是在这一意义上，建造处理自然语言的计算机系统跟建造其他大型的计算机系统一样，主要是一种工程性的工作。跟其他系统建造工作一样，计算语言学采用模块

① 详见石纯一等（1993），第 355—363 页。

② 参考 Halvorsen（1988），pp. 202—203.

化(modularity)和建立形式模型(formal models)两种通用技术。所谓模块化指把我们的系统所涉及的知识分割为相对独立的成分,然后分别攻克一个个子问题,从而缩小整个系统的规模。所谓建立形式模型指为复杂系统建立一种相对简单的抽象模型,然后为这种简化的模型设计我们的计算机系统。[①]

这种工程主义取向的计算语言学研究是有很强的应用动机的。因为语言是人类交际和记录信息的工具(vehicle),如果使计算机获得生成和理解自然语言的能力,那么计算机就能执行只有人类才能完成的工作,诸如翻译、文本处理、信息抽取和检索等;所以,能处理自然语言的计算机系统将使计算机更为有用。[②] 也就是说,通过计算语言学的研究,可以开拓更多的计算机应用领域。

2 工具主义取向:着眼于用计算机作语言分析

计算语言学最宽泛的定义是:用计算机来研究和处理自然语言。例如:

(1) 计算语言学是关于自然语言的计算机处理的一门学科。它用计算机技术来研究和处理自然语言。

——陆致极(1990),第15页

(2) 计算语言学是利用电子数字计算机进行的语言分析。……计算分析最常用于处理基本的语言数据——例如建立语音、词、词元素的搭配以及统计它们的频率。

——《大不列颠百科全书》,转引自翁富良等(1998),第1页

(3) 对计算语言学一般有狭义的和广义的两种理解。狭义理解盛行于计算语言学最为发达的美国,它大致上就是人工智能中自然语言理解(包括机器翻译)的理论和方法部门,它的操

① 详见 Grishman (1986), pp. 7—8.

② 详见 Grishman (1986), p. 1.

作内容大致上就是上面所提到的(1)—(5)。① 广义的理解则把凡是利用计算机处理自然语言的有关问题(例如,风格研究)都囊括进来了,这种理解欧洲比较盛行。

——钱锋(1990),第27—28页

在这种包容性很大的定义中,除了有§1中讨论的研究能理解自然语言的计算机系统之外,还有利用计算机来进行跟语言相关的研究等内容;比如,用计算机对字母频率、汉字频率、词长、句长、句型等语言成分的统计研究,以及建立在语言成分的统计基础上的作品风格研究和匿名作品的作者考证研究等。简单地说,工具主义取向的计算语言学着眼于用计算机来进行语言的计量研究(quantitative studies)。

值得一提的是,随着用计算机来采集、整理、加工和管理语言材料工作的深入开展,逐步形成了语料库语言学(corpus linguistics)这门计算语言学的分支学科。大概地说,语料库语言学研究机器可读的(machine-readable)自然语言文本的采集、存储、检索、统计、语法标注(grammatical tagging)、句法语义分析,以及具有上述功能的语料库在语言定量分析、作品风格和作者考证研究、词典编纂、自然语言理解和机器翻译等领域中的运用。比如,为了研究现代美国英语,美国的布朗大学在1964年建立了库容量为100万词的Brown语料库。为了研究现代英国英语,英国的兰开斯特大学跟挪威的奥斯陆大学、卑尔根大学在20世纪70年代合作建成LOB语料库,库容量也是100万词。欧美各国的学者利用这两个语料库开展了大规模的英语研究。在1970—1978年间,他们用86种词类标记对布朗语料库进行语法标注。Greene和Rubin还设计了名叫TAGGIT的自动标注系统,其庞大的规则库里有3300条上下文有关规则。TAGGIT系统对布朗语料库的全部100万词语料进行自动标注的正确率达77%,其余的同形和兼类歧义问题最后由人工来解决。②

从方法论上看,语料库语言学跟工程主义的计算语言学很不相

---

① 这里的(1)—(5)就是§1中第一～三步的内容。

② 参考黄昌宁(1990),第43—44页;冯志伟(1992),第90页。

同。后者采用的是以知识(表示成规则)为基础的方法,即人工智能的方法。这种方法假定:如果计算机要处理自然语言,那么它必须跟人一样具有句法、语义、语用、话语篇章、主题事物、周围世界等方面的知识和逻辑推理能力。因为人处理语言时的心理状态和心理过程就是这样的,计算机必须具有跟人相同和相近的知识才能处理自然语言。而语料库语言学采用的则是以语料统计为基础的方法,即基于概率的方法。这种方法认为:计算机并不能像人一样利用知识去理解语言,人们也无法把理解语言所需的各种知识形式化地表示成规则。有鉴于此,这种方法假定:如果我们能对数量很大的语言数据作出定量化的统计分析,那么我们就能对语言成分的分布和语言成分之间的关系等进行概率性的预测,从而补偿计算机缺乏知识和推理能力的缺点。[①] 比如,在 1978—1983 年间,英国的 Leech、Sampson、Garside 等人对 LOB 语料库进行词类标注实验。为此,他们还设计了一个名叫 CLAWS 的系统(Constituent-Likelihood Automatic Word-tagging System)。他们完全放弃了传统的规则模型,把自动标注的算法建立在统计信息的基础上。他们采用了 133 种词类标记,利用已带有语法标记的 Brown 语料库来获取两个相邻标记的同现频率,据此建立了一个规模为 133×133 的"标记转移概率矩阵"(tagging transition probability matrix),用以反映在前一种标记的条件下后一种标记出现的概率。整个语法标注过程所依据的知识都是由这个矩阵提供的。CLAWS 系统对 LOB 语料库的全部 100 万词语料进行自动标注的正确率达 96%,比基于规则的 TAGGIT 系统提高了将近 20%。[②] 例如,对于句子"Henry likes stews.",其中 Henry 是名词短语,只有 NP 一种标记;likes 和 stews 可以是名词复数和动词第三人称单数,因而有 NNS 和 VBZ 两种标记。于是,这三个词可以有下列四种词类搭配方式:

i. NP+NNS+NNS+.=17×5×135=11475

ii. NP+NNS+VBZ+.=17×1×37=629

---

① 参考桂诗春、宁春岩(1997),第 138—149 页。

② 参考黄昌宁(1990),第 44 页;桂诗春、宁春岩(1997),第 145 页。

iii. NP＋VBZ＋NNS＋. ＝7×28×135＝26460

iv. NP＋VBZ＋VBZ＋. ＝7×0×37＝0

在这些由形式类表示的搭配方式的右侧（等号后面）给出每种标记跟相邻标记的同现概率，并用这种概率的乘积作为决定某种搭配方式的概率的变量。假定决定某种搭配方式的概率等于该变量除以所有变量的和，那么第三种搭配的概率最高（26460/11475＋629＋26460＋0＝69％）。系统可以据此确定句子“Henry likes stews.”的形式类标记是 NP＋VBZ＋NNS。[①] 既然通过概率计算可以确定兼类词在某种组合中的词类属性，那么由兼类词引起的结构歧义也可以通过概率计算来消歧（disambiguation 或 ambiguity resolution）。于是，基于语料库的统计模型不仅可以用来解决自然语言的语法标注任务，而且还可以运用到句法、语义等更高层次的分析上来。[②]

3 认知主义取向：着眼于人类使用语言时的心理过程

在计算语言学的定义中，为数不多的涉及人类使用语言时的心理过程。例如：

(1) 计算语言学是一门计算机科学和语言学紧密结合的科学。它用数学的方法来制订语言规则和模型去解决有关计算机的语言学习和理解、语言信息的存储、组织、更新、转换和生成等问题。在这些问题中，核心是学习和理解。

——黄建烁（1991），第 24 页

(2) Computational linguistics is best viewed as branch of artificial intelligence (AI). As all fields within AI, it is concerned with the investigation and modeling of a cognitive capacity. In the case of computational linguistics it is the lan-

① 参考桂诗春、宁春岩（1997），第 138—149 页。

② 参考黄昌宁（1990），第 44 页。

guage capacity that it is in focus. However, the concern is not necessarily to construct a psychologically realistic model of human behavior. The goal is rather to identify and characterize the classes of processes and the types of knowledge which are implied by the ability to communicate and assimilate information using natural language regardless of their psychological status.

——Halvorsen (1988), p. 202

（计算语言学最好看作是人工智能的一个分支。跟人工智能的所有其他领域一样，它涉及对认知能力的研究和建模。在计算语言学这里，它着重的是语言能力。但是，这种研究不必去建构关于人类行为的具有心理真实性的模型。其目的就在于确定和刻画用自然语言进行交际和获取信息的能力中所包含的知识的种类及相关处理过程的类别，而不管其实际的心理状态。）

黄建烁(1991)的定义为计算语言学确立了一种非常宏伟的目标，那就是教会机器自动地学习，即让机器理解语言并自动地学习和更新知识。用 Hans Karlgreen 教授的话来说，就是“用计算的方法来制定人类语言行为的模型，并以此去了解人们怎样听说读写，怎样学习新知识和更新旧知识，又是怎样理解、存储和组织语言信息的”。他甚至认为，计算语言学的一个最根本的问题就是了解“人类的大部分活动在什么程度上能够简化成机械的操作”。[①] Halvorsen (1988)则强调，计算语言学是对人类语言处理能力和心理过程的功能（而不是结构）模拟。这就是典型的人工智能方法。这种功能模拟的方法直接影响和促成了认知心理学的基本信念：可以把计算机作为人类思维的模型，也可以用计算机来模拟人类的认知过程。

T. Winograd (1983) *Language as a Cognitive Process*(〔把〕语言作为一种认知过程〔看待〕)，则可以说是认知主义取向的杰出典范。他由下列两个问题激发灵感，尝试建立一种语言研究的认知范

① 详见黄建烁(1991)，第 31 页。

式(cognitive paradigm):

i. 一个人要说话和理解语言,必须具有哪些知识?

ii. 为了在交际中使用这些知识,人的心智是怎样组织的?

他把语言使用看作是一种以知识为基础的交际过程,认为人无论是说话还是听话都必须具有一定的知识;比如,词序规则、词汇和词的结构、语义特征、所指关系、时制系统、话语结构、说话人的态度、韵律规约、风格规约、世界知识等。在理论方面,他企图探讨人是怎样习得、运用这些知识的;在实际运用方面,他尝试用计算机来模拟人习得、储存、运用这些知识的过程,所以他又称这种范式为计算的范式(computational paradigm)。[①]

这种取向的学者喜欢用认知心理学的眼光来看待语言使用。从信息加工过程的观点看,人说出一句话和理解一句话时,在大脑中有一个关于所描述的外部世界中的事物或事件的心理映象,可以称之为内部语言;而人处理语言的过程就是把外部语言转化为内部语言,经过加工后再由内部语言转化为外部语言的过程。计算机也可以用类似的过程来处理自然语言:首先确定一种语言的内部表示;然后,寻求一种把所限定的语言子集中的语句转换为内部表示的方法。在他们看来,让计算机理解语言的关键是:应能对一般的自然语言的句子作出语义解释,即设计一种一般的内部表示。内部表示是自然语言处理的关键,它影响着系统对语言知识和世界知识的描述和利用,因此也影响着整个处理系统。[②]

不同的学者由于对人类处理语言的心理过程的认识不同,因而采用了不同的理论和方法来建造自然语言处理系统。其中,一类系统比较重视句法分析,尽管所依据的语法理论各不相同。比如,Winograd 于 1972 年研制了关于积木世界的 SHRDLU 系统;该系统可以接受命令,通过一只机械手对积木进行操作,回答有关积木世界所处的状态的问题。他认为句法需要解决的问题是:语言究竟是

---

① 详见 Winograd (1983), pp. 1—34. 另外,参考黄奕(1985)对该书的介绍和评论。

② 详见杨抒(1988),第 21—23 页。

怎样组织起来表达语义的？他采用 Halliday(1967、1970)的系统语法(Systemic Grammar)，把句法结构看作是生成句子的过程中一系列句法结构选择的结果。比如，小句(clause)是由主句性(major)和从句性(secondary)两个特征构成的一个系统，它们是互相排斥的句法特征，任何小句只能选择其中的一个特征；陈述(declarative)、祈使(imperative)、疑问(question)三种句法特征构成一个系统，主句(major clause)必须选择其中的一种句法特征。语义根据一定的外部世界模型作出推论来指示句法分析，从而得出句子的正确的语义解释。例如，在“I rode down the street in a car.”中，只有运用世界知识(街道不可能在汽车里)作出推论，才能排除 in a car 作 street 的修饰语。SHRDLU 系统的工作过程是：运用扩充转移网络(augmented transition network, ATN)的句法知识和跟积木世界相关的语义知识对输入句进行句法语义分析，然后直接运用过程表示的(procedural)知识对输入句进行推理，最终找到所需执行的过程知识并执行之。Woods 于 1972 年设计了关于月球化学成分的 LUNAR 系统，该系统的句法部分根据 Chomsky(1965)的转换生成语法，分析出标准理论所指定的深层结构，再输入语义部分。语义部分根据句法上的深层结构再进行语义信息的分析。数据检索部分再根据输入句的语义编译成一种面向系统的形式语言(即查询语句)，以便直接查询数据库，并最终产生结果(即回答)。Simmon (1973)根据 Fillmore (1968)的格语法(Case Grammar)建立了语义网络理论。他采用 Woods 的 ATN 来分解输入句的句法关系，同时分析深层格结构，记录语义关系；最后求出输入句的语义关系，据此来理解语义。另一类系统不作详细的句法分析，直接从语句中抽取语义信息。比如，Yorick A. Wilks 认为，整段言谈的内容是由一些简单的基本信息构成的。一个复杂的句子也是由基本信息通过概念连结成实时的线性序列，而不是语言学家所认为的具有层次的树形结构。在这种思想指导下，Wilks (1973)用人工智能的方法设计了一个英法机器翻译的模型。英语词汇量 600，用以组成英语日常用语；把简单的段落输入计算机，能译成通顺的法语输出。这个模型不作句法分析，而是用一套“语义模板”来接受输入句中的信息。也就是说，该系统把

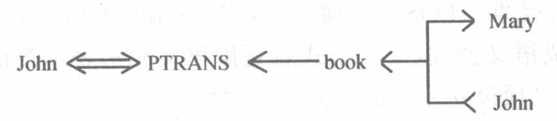
源语言的输入语句直接处理为一种语义结构，作为一种中介成分，再据此生成目标语言的语句，也可以在这种中介成分上作谓词演算用于特定领域。语义结构分为三层：模板（templates）、公式（formulas）、元素（elements）。其中，元素是基本的语义单位，包括语义特征（如：MAN、THING、FORCE、CAUSE）和语义格（如：SUBJ〔施事〕、OBJE〔受事〕、POSS〔领属〕、SOUR〔来源〕、GOAL〔目标〕）。语义元素构成语义公式，语义公式就是用语义元素表示英语的词汇意义；每一个义项设立一个语义公式。比如，

interrogate：((MAN SUBJ) ((MAN OBJE) (TELL FORCE)))

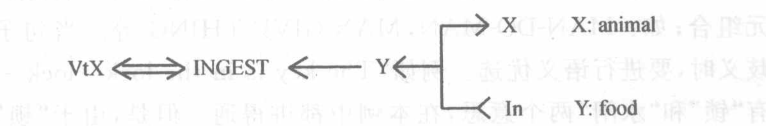
审问：人迫使人说话

语义公式构成语义模板，语义模板是由"施事-动作-受事"构成的三元组合；如：MAN-DO-MAN，MAN-GIVE-THING 等。当句子有歧义时，要进行语义优选。例如"The key is in the lock."lock 一词有"锁"和"水闸"两个意思，在本例中都讲得通。但是，由于"锁"和"钥匙"语义联系的程度高，因而优选"锁"。对于某些歧义的情形，还需要运用常识推理才能作出判断。Roger C. Schank 认为人脑中存在着某种概念基础（conceptual base），语言理解的过程就是把语句映射到概念基础上去的过程。概念基础具有完善的结构，人往往能根据初始的输入预期可能的后续信息。句法分析对语言理解的用处不大，因为语言理解需要的是输入句的意思，而不是它的句法结构。计算机要理解语言，必须模拟人的心理过程；要像人一样根据上下文、环境、知识、记忆等作出预期（expectation），从而获取语义。句法只起一种指引的作用，即根据某些输入词语形成概念结构，预期它的句法形式，便于查找核实。Schank (1973)提出了概念从属（Conceptual Dependency，CD）理论，建立了 MARIE 模型。这个模型用同义互释（paraphrase）的方式来检验计算机对自然语言的理解程度。即输入一句话，要求计算机用另外的一些语句来解释。CD 理论提供一组原始行为（primitive act）和一组概念及其相互之间的从属关系，作为构造 CD 表达式的基础。CD 中的原始行为有：ATRANS（抽象

关系的转移，如 give)，PTRANS(物理位置的转移，如 go)等十几个；有六种概念范畴：现实世界中的事物 PP 及其属性 PA、动作 ACT 及其属性 AA、还有时间和方位。概念从属关系有：PP⇔ACT(施事与动作之间的施动关系)，PP↔PA(当事与属性之间的描述关系)等。比如“John gave Mary a book.”的 CD 表达式(有删略)是：



每个动词按照从属关系编入动词词典。语句输入，根据句法指引找出主要的名词和动词。再查动词的概念从属关系，联系句中的名词获得语义解释。比如，输入“John ate the steak.”，查 eat 条的注解(有删略)为：



代入句中名词就得到这个句子的 CD 表达式(有删略)：

John ⇔ PTRANS ← book ← Mary / John

上述这些不同的理论和方法，都是基于研究者对于“人是怎样理解语言的”这一问题的不同见解而发展出来的。也就是说，他们分别用不同的计算范式来实现其认知范式。[①]

4　实证主义取向：着眼于检验语法理论的可靠性

跟 § 1 所述的抱有实用目的的工程主义取向不同，大多数计算

① 详见杨抒(1988)，第 22—26 页；范继淹、徐志敏(1980)，第 9—19 页。

语言学研究并不跟某种特定的应用目标相挂钩，而是另有某种科学研究的目标。其中之一就是用计算机来对语言学家提出的各种语言学理论进行检验。比如：

> One natural function for computational linguistics would be the testing of grammars proposed by theoretical linguists. ——Grishman (1986) §1.1, p.5
>
> （计算语言学的一个自然的功能是对理论语言学家提出的各种语法进行检验。）

用计算机来检验某种语法理论或某组语法规则，这对语言学家来说实在是一件既令人兴奋又令人不安的事。兴奋的是语言学的理论和规则居然可以像数学公式一样让计算机去执行，不安的是能顺利通过机器检验的希望是极其渺茫的。Friedman (1971)还真的设计了一个检验转换生成语法的系统，名叫 Friedman's Transformational Grammar Tester。该系统可以按照转换生成语法来生成句子，于是语言学家可以用它来检验他们的语法是不是真的只生成合语法的句子。事实上，由于大多数语言学理论的形式框架（包括：移位规则的性质、对转换的限制、语义解释规则的形式，等等）都是有问题的，而且理论语言学的重点并不是建造一种能适应计算测试的实体性的语法；因而就目前来看，作为语言学理论的测试工具，计算机的用处是不大的。①

看来，让计算语言学来充当语言学理论的审判官是不合适的；它会导致两种消极的后果：(i) 计算语言学对理论语言学的失望和抱怨，漠视理论语言学的研究成果，撇开语言学家的工作另搞一套；(ii) 理论语言学对计算语言学的敌视和疏远，拒绝采用计算机科学的理论、概念和方法来研究语言，使语言学研究失去一种丰富的理论营养和强劲的应用动力。更为现实的定位是：把计算语言学看作理论语言学和计算机技术的桥梁，通过计算语言学家的工作来沟通语言学理论和计算机技术，来形成语言学技术（linguistic technology,

---

① 详见 Grishman (1986), p.5.

如：针对某种语法体系的语法解释器和分析器，言语合成算法等），从而完成语言学理论在计算机上的应用。因为，在语言学理论和计算机处理技术之间存在着很深的鸿沟，一般的语言学理论研究的是抽象的语言能力(competence)，即理想的说话人和听话人的内在的语言知识；而不研究具体的语言运用(performance)，即语言知识在实际的语言活动中是怎样运用的。但是，计算机只能处理活动和过程性的知识。因此，计算语言学一直在尝试通过把语言学理论转变为算法(它能模拟遵守语言学理论和语言能力语法中所包含的各种语言学限制和概括的语言行为)，来沟通语言能力语法和某种要适应用机器处理的特定的语言运用。[①] 事实上，更大的矛盾在于：语言学理论基本上是描述性的，而计算机技术中的算法描述和编程语言则基本上是过程性的。下面，我们简要地讨论这种矛盾及其解决办法。

一般地说，计算机要处理自然语言(最终目的是抓住句子的意义)，首先必须对输入句进行句法分析(parsing)，从没有显性结构标记的符号串上找出结构来；即识别输入句的各个构成成分以及它们之间的关系，比如确定句子的主要动词及其主语和宾语，确定修饰成分及其中心语等。要分析句子的结构就需要语法的指导，正是语法提供了一种语言的结构成分和符号串跟结构之间的关系的明确定义。在计算语言学上，通常称一个能根据一部特定的语法来分析句子(确定句子的推导过程)的程序为分析器(parser)。这种分析程序主要涉及两部分内容：(i) 一组语法规则，它们由某种形式化的语法理论组织在一起，形成某种语法形式体系(grammatical formalism)；(ii) 一种控制机制(control mechanism)，它决定在分析过程中怎样运用语法规则、怎样保持对于各种业已发现的成分的记录、使程序在有限步运算后找出结构，即形成某种分析算法(parsing algorithms)。大家知道，程序是用编程语言(programming languages)编写的。而编程语言基本上是过程性的表示体系(procedural representation)，因为编程的目的本来就是给计算机提供一套明确而详尽的怎样干某事的指令(instructions)。但是，语法规则通常都是陈述性的(declar-

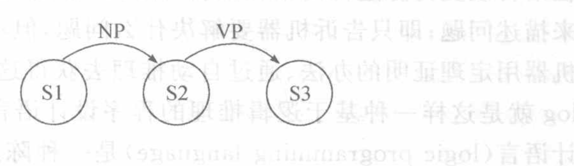
① 详见 Halvorsen (1988)，pp. 200—201.

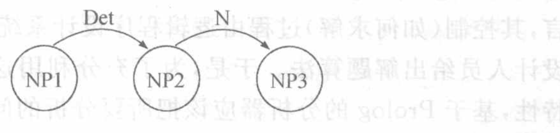
ative)，而不是过程性的；它可以告诉我们一个句子往往由一个 NP 和一个 VP 构成，但并不告诉我们怎样用一个 NP 和一个 VP 去构成一个句子。面对这种语言学理论和计算机技术之间的不适配，有两种解决问题的思路：第一种，把陈述性的语法形式体系改变为过程性的语法形式体系，用过程性的形式体系来表示和组织语法规则。比如，利用转移网络这种形式机制的 RTN 语法（recursive transition network grammar）和 ATN 语法就是一种过程性的语法体系。例如，下面是一部小型的上下文无关语法的三条规则及相应的三个转移网络：

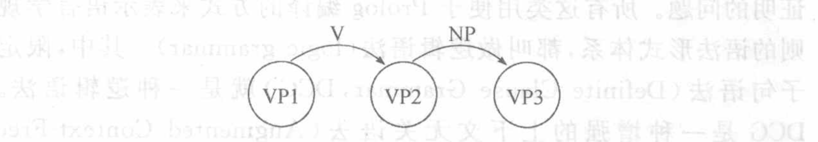
S→NP＋VP

NP→Det＋N

VP→V＋NP







可见，RTN 由一组子网络组成。每个子网络由一组状态构成，表示一种句法范畴（如 S、NP、VP 等，其后的数字是该范畴的状态编号）；不同的状态之间用带箭头的弧线连结，弧线上面标明该范畴的组成成分的句法范畴（如 Det、N 等）。在 RTN 中，任何一个子网络都可以调用包括自己在内的任何子网络。比如，上图中 S 子网络要调用 NP、VP 子网络，VP 子网络要调用 NP 子网络。RTN 基本上等价于一部上下文无关语法（即短语结构语法）。

一个 RTN 可以看作是一组地图，它们指引着你去发现句子是怎样由短语构成的、短语是怎样由词构成的。因此，RTN 语法是一种过程性的形式体系。ATN（augmented transition network）则是对 RTN 的扩充，它通过增加一组寄存器来储存分析过程中得到的中间结果（如局部句法树）和有关信息（如名词短语的人称和数、某些语言成分的语义特征等），还在每条弧上附加任意的条件测试（符合条件，即测试成功才能通过这条弧）和动作（当通过一条弧时，相应的动作便依次执行；这些动作主要用来设置和修改寄存器的内容）。正是这种条件和动作，使得 ATN 这种表示方式从语法形式体系转变为一种分析算法。但是，这种扩充破坏了语法形式体系的陈述性的本质，同时用 ATN 的分析器在控制策略的使用上备受限制；因此，近年来这种分析器已经不太时兴了。第二种思路是，把过程性的编程语言改变为陈述性的编程语言，用陈述性的表示体系（逻辑形式）来描述问题；即只告诉机器要解决什么问题，但不说怎样去解决，让机器用定理证明的办法、通过自动推理去获得这方面的信息。Prolog 就是这样一种基于逻辑推理的程序设计语言，这种逻辑程序设计语言（logic programming language）是一种陈述性（表示问题）语言，其控制（如何求解）过程由逻辑程序设计系统本身实现，无须程序设计人员给出解题算法。于是，为了充分利用这种编程语言的内在特性，基于 Prolog 的分析器应该把所要分析的问题看作是一个定理证明的问题。所有这类用便于 Prolog 编译的方式来表示语言学规则的语法形式体系，都叫做逻辑语法（logic grammar）。其中，限定子句语法（Definite Clause Grammar，DCG）就是一种逻辑语法。DCG 是一种增强的上下文无关语法（Augmented Context-Free Grammar），它的生成能力不低于 ATN 语法。更为重要的是，用限定子句表示的语法规则本身就是逻辑程序设计语言 Prolog 的可执行程序。换句话说，Prolog 系统可以直接解释用 DCG 形式表示的语法规则，而无需像 ATN 那样另外再设计一个句法分析器（规则解释程序）来完成这个任务。下面，我们来看一部简单的上下文无关语法是怎样用 DCG 这种形式体系来描述的：

sentence→noun-phrase，verb-phrase

noun-phrase→determiner, noun

verb-phrase→trans-verb, noun-phrase

determiner→[the]

noun→[man]

noun→[wine]

trans-verb→[likes]

Prolog 系统可以把这样书写的 DCG 规则直接翻译成 Prolog 可执行程序，例如：

```
sentence (X, Y): ——noun-phrase (X, Z),
                   verb-phrase (Z, Y).
noun-phrase (X, Y): ——determiner (X, A),
                      noun (A, Y).
verb-phrase (X, Y): ——trans-verb (X, A),
                      noun-phrase (A, Y).
determiner ([the|X], X).
noun ([man|X], X).
noun ([wine|X], X).
trans-verb ([likes|X], X).
```

在 Prolog 的规则中，每个非终结符都被改写为具有两个变元的复合项。其中，每个变元都是一张表，并且第二个变元是第一个变元的余表。例如，输入句子"The man likes wine."可表示为目标：

```
? —sentence ([the, man, likes, wine], [ ]).
   yes
? —noun-phrase ([the, man, likes, wine], X).
   X=[likes, wine]
```

第一个问题问词串[the, man, likes, wine]是不是一个句子，机器回答是；第二个问题问在同一词串中扣除什么就是一个名词短语，机器回答扣除余串[likes, wine]。可见，计算机技术和语言学理论是相互影响、相互促进的。这造成了计算语言学和理论语言学的紧密合

作，并且产生出丰硕的成果。比如，广义短语结构语法（Generalized Phrase Structure Grammar，GPSG）和词汇功能语法（Lexical Functional Grammar，LFG）都是陈述性的语法形式体系，它们都受到 M. Kay（1979）的计算语言学著作 *Unification Grammar*（合一语法）的影响。其中，LFG 是理论语言学家（J. Bresnan）和计算语言学家（R. Kaplan）的合作成果，GPSG 的部分作者担任过大型的计算语言学项目的顾问。随着这种理论语言学和计算语言学的会聚（convergence），也有许多计算语言学项目采用 GPSG 或 LFG 作为其语法形式体系，从而实现了从语言学理论到计算机技术的转变。[①]

5 逻辑主义取向：着眼于语言学知识的自动发现

值得注意的是，最近出版的一些计算语言学著作，作者在计算语言学的定义中特意强调了语言的计算结构和计算模型。例如：

> （1）计算语言学旨在以自然语言处理（包括理解、生成、人机对话、机器翻译以及语音/文字输入的后处理等）为技术背景，揭示自然语言的词法、句法、语义、语用诸平面及其相互作用的计算结构，把语言学知识重塑成可以转化为产品的计算模型。
>
> ——白硕（1995），第 2 页
>
> （2）现代计算语言学是通过建立形式化的计算模型来分析、理解和处理语言的学科。……广义地讲，计算语言学是研究字符串的结构以及结构和意义的关系的学科。
>
> ——翁富良、王野翊（1998），第 1、9 页。

按照白硕（1995）的理解，要建造一个处理自然语言的计算机系统，必须有大量的语言学知识作后盾；但是，语言学知识的发现工作主要是以手工的方式进行的。因此，利用计算机来自动（或辅助）发

① 以上内容详见 Halvorsen（1988），pp. 204—210；Gazdar & Mellish（1987），pp. 228—229，pp. 229—235；石纯一 等（1993），第 64—68 页；第 333—422 页。

现语言学知识，将极大地提高研究的效率、扩大研究的规模、把语言学家从收例句、制卡片、画表格等烦琐的事务中解放出来。可见，研究语言学知识的计算机辅助发现系统，是计算语言学的一个别开生面的研究方向；这种工作不仅有助于我们以计算机为模型来研究儿童语言习得，而且对于开发自然语言处理系统也具有实用的价值——一个语言学知识的计算机辅助发现工具实际上相当于一个使自然语言处理系统具有自扩充、自维护功能的高级开发工具。所谓语言学知识的发现，指的是从一个由例句组成的语料库中发现特定的自然语言规律。这种从一组事例中发现一般规律的认知活动，在逻辑上被描述成一种"归纳"过程。但是，历来对于归纳的研究跟逻辑是脱节的，特别是对于语言学知识的发现的逻辑实质的研究是十分缺少的。作者决心研究语言学规则这种特殊形式的知识的发现的逻辑实质，全面地展示跟语言学知识发现有关的各个层次上的形式化机制——从数学建模、逻辑分析、算法描述、具体实现直到结果的语言学解释。作者采用语言学中经典的分布分析的思想，并针对真实语料的各种特点，结合汉语的实际，从数学、逻辑、算法以及实现各个角度，全面阐述了从语料中发现确定性语言学知识（主要是词类和句法规则）的理论和方法。作者首先从数学角度讨论了分布理论的完善和推广，分别在词、短语、词结（word complex，即超距相关的实词多元组，long-distance dependent word n-tupple，如："英语我十年前就会说了"中的"英语……说"）的划类问题上引入分布分析方法。作者在讨论词类及其划分的数学理论时，提出了词类划分的不动点理论、指出分布分析的任务是求解最大不动点，澄清了语言学界有关分布分析中含有"逻辑循环"的误解、证明了最大不动点在极限意义下的可计算本性、明确了分布分析方法的两个基本的逻辑前提：词的同一性和语言边界的明确性，从而解决在词类问题上"发现什么"和"能否发现"两大问题。在讨论发现句法规则的数学理论时，作者用构造性的方法建立一个基于句型推衍的变换规则系统，用以说明什么是基本句型和怎样从一些句型得到另外一些句型；其中，推衍规则包括句型推衍规则和环境推衍规则，它们都是重写规则（rewrite rules）；并阐明这种规则发现系统跟分布分析的关系：同分布关系和

作为重写规则的推衍规则在本质上都是一种“替换”。就这样，作者从词的分布分析推广到了短语结构的分布分析，接下来他又把分布分析推广到词结。作者发现如果两个词结是同分布的，那么它们一定同时满足或不满足任何一个变换；所以变换是实词多元组和多元句法环境之间的一种推衍关系，词结是变换下的不变量、多元环境的填充物，而多元环境则是某一句法结构中抠掉了词结的剩余部分；由于词结是以各种不同的多元环境作为分布框架的，因而变换分析就是词结的分布分析，通过变换分析可以给词结进行分类。这样，句子可以看作是由词结加上环境构成的，句子语义恰好可以分解为词结的语义加上环境的语义。比如：

“河不过了”，指的是撤销“过河”的意愿；

“饭不吃了”，指的是撤销“吃饭”的意愿。

多元环境“不……了”的语义为“实现事件 E 的愿望撤销了”，加上由词结“过……河、吃……饭”的意义正好是句子的意义。作者甚至希望通过词结的分布分析，来归纳词结中的从属成分的语义格；其根据是词结的同分布类跟内部语义角色关系和外部组合能力相同的语义结构类是大致对应的，这样，同分布的词结的相同位置上的从属成分的语义格是相同的；比如，上例中“河、饭”的语义格是一致的。接着，作者讨论了语言学知识发现的逻辑基础，提出了一个进行逻辑聚类的类似缺省逻辑（default logic）的非单调形式演算系统，用以解决在分布知识不完全的情况下进行分布分析的逻辑聚类方法及其逻辑合理性问题。在此基础上，作者提出了语言学知识发现的两种实现算法：交互式聚类和无反例聚类。前者是增量式的，符合语言学家的工作习惯——提出各种正例和反例来发现区别，细化一个业已存在的规则系统；后者是批量式的，符合语言工程师的工作习惯——把一个没有反例的大语料库交给计算机去运行，中间不作任何干预，只管到时候取结果。从而，在技术上解决了词类等语言学知识如何发现的问题。最后，作者用交互式算法建立了一个实验系统 CASD—1，算是对上述理论的实践或检验。这是一个面向汉语的词类划分系统，它通过对例句文件中具有代表性的 66 个例句（47 个正例、19 个

反例）的分析，通过变换和把体词性成分抽象为词类范畴来构造句式，再通过人机界面请人判断正误以获取分布信息，结果该系统分出动词的15个分布上的小类；比如，"送、嫁、卖"是一类，"炒、织、准备"是一类。接着，作者考察这15个小类的语言学意义，发现每一类动词都有独特的句法分布（比如，能否进入双宾句式、能否后附"着"）、语义上它们有独特的格角色结构（论元结构，比如，其主语位置上的NP是施事还是处所、其宾语位置上的NP是受事还是结果）、语用上它们隐含独特的预设集合（比如，"送"类动词表示的给予事件的预设是：事件发生前，受事为施事所有；事件发生后，受事为与事所有）。这种计算语言学工作对语言学家来说是比较亲切的，因为它在相当程度上模拟了语言学家发现语言学规则的过程。

白硕（1995）的研究有着明显的逻辑主义追求，那就是通过研究语言学知识的发现来探索归纳法的逻辑机制和计算结构。一般地说，从逻辑上看，人类的思维活动不外基于演绎法和基于归纳法两类。演绎法常常是从一些多少已经抽象化、形式化的前提出发，演绎出种种结论来。只要前提中含有可以互相消解（resolve）的对象，就一定可以衍生出新的命题来。显然，从前提演绎出结论是计算机可以胜任的工作。而归纳法常常是从未充分抽象化、形式化的大量个别事例出发，希望从中抽象出有用的概念、模式、定理来。这种工作能不能用计算机来完成呢？由于在使用归纳法的时候（比如，划分词类、发现句法模式等），目标的确立、是否达到目标的判别、达到目标的手段的建立等都是通过反复尝试而逐步建立起来的。对于这种缺少确定性的过程，计算机是很难单独完成的。怎么办呢？答案是建立一个人机共生的系统，由人来负责设定目标和手段、由机来负责实现这种手段而不管目标是什么。如果有了这样的人机共生系统，就可以大大地提高工作的效率和质量。要想做到这一点，就必须进一步研究归纳的手段和逻辑机理。而白硕（1995）主要是以语言学问题为背景，提出许多关于归纳的概念和方法作为人机共生系统的基础。[①] 他特别强调归纳的非单调性、可错性的特点：已经归纳出来的

---

① 详见马希文为白硕（1995）所写的序，第ii—iii页。

规则总有可能被后来的事实证明是不正确的、需要修改的，然而在没有遇到这样的事实时，这些规则又可以认为是近似正确的、不妨使用的。作者就是用这种允许某种"逻辑跳跃"来达到一些好的猜测的方法来发现词类和句法规则，并希望这种机制不仅仅局限于语言学知识的发现，希望这种研究对于探索知识发现的一般途径、对于认识归纳和类比的逻辑实质有所贡献。

从方法论和哲学背景上看，计算语言学研究有理性主义和经验主义两大分野。理性主义方法认为：人的很大一部分语言知识是与生俱来的，即是由遗传决定的。受Chomsky内在语言官能(innate language faculty)学说的影响，计算语言学界很多人信奉理性主义。他们秉承人工智能研究中的符号主义传统，通过人工汇编初始语言知识(主要表示成形式规则)和推理系统来建立处理自然语言的符号系统。这种系统通常根据一套规则或程序，将自然语言"理解"为某种符号结构；再通过某种规则，从组成该结构的符号的意义上推导出该结构的意义。在一个典型的自然语言处理系统中，句法分析器(parser)按照人所设定的自然语言的语法把输入句分析为句法结构(一种特定形式的符号结构)，再根据一套语义规则把语法符号结构映射到语义符号结构(如：逻辑表示、语义网络、中间语言等)。由于自然语言处理系统中的规则集通常是先验的，即是由人设计好以后赋予机器的；因而，这是一种典型的理性主义的方法。经验主义方法认为：人的知识只有通过感官传入、再通过一些简单的联想(association)和泛化(generalization)的操作才能获得，人不可能天生拥有一套有关语言的原则和处理方法。表现在计算语言学中，许多研究尝试从大量的语言数据中获取语言的结构知识，从而开辟了基于语料库的计算语言学这种经验主义的研究方法。其中的神经网络方法秉承了人工智能研究中的连结主义传统，由机器通过学习给定的实例(训练数据)之间的输入-输出关系，来获得神经元(人工神经节点)之间的连结强度(strength，或称"权"weight)，以反映从输入状态到输出状态之间的映射关系。其中的统计学方法试图建立统计性的语言处理模型，并由语料库中的训练数据来估计统计模型中的参数。比如，§2中介绍的词类的自动标注，其做法是先使用少量已经人工标

注的语料进行训练，然后将学到的词类标记的共现概率分布用于标注尚未标注的文本。这都是通过学习训练实例来获得某种语言处理能力的，因而是典型的经验主义的研究方法。[①] 简而言之，理性主义强调基于规则的方法，经验主义强调基于学习的方法。而白硕(1995)的工作则尝试兼采这两种方法之长又避免这两种方法之短。粗略地说，这是一种企图发现规则而不是赋予规则、基于语料库但不拘于统计学方法的路子。作者考虑到仅靠统计学方法是无法从语料中发现确定性的语言学规则的，因而尝试一种从精炼语料库中动态地归纳规则的方法。这种从语料库中通过学习来获得符号处理系统中的规则集的方法，在本质上是归纳逻辑。这种方法一方面用到符号处理系统中的规则表达，但规则又是从语料库中经验地获得的；因而，就其本性而言是一种经验主义的方法。[②]

从语言学的角度看，白硕(1995)给人印象最深的是：对分布这一概念的实质的揭示、对分布分析方法的全面推广。白硕(1995，第111—112页)指出，分布是一个十分深刻的语言学概念，分布概念的实质就是某些语言学对象在特定含义下的可替换性。有了可替换性，这些语法功能一致的语言学对象才能聚集成类，才有可能总结出能概括普遍语言现象的规则。因此，在句法范畴和句法规则的发现过程中，分布分析的方法起着核心的、决定性的作用。从数学上看，分布分析的实质是等价类划分。要进行等价类划分，就要定义相应的等价关系，即所谓的同分布关系。同分布关系的定义取决于环境的定义，而定义环境又需要有一个初始的等价类划分。于是就需要一整套的不动点理论使分布分析走出“逻辑循环”的陷阱。分布分析不仅能发现句法范畴，而且能发现句法规则。由于句法规则是一种重写规则，而重写就是替换；在这里，被替换的恰恰是短语结构。因而，正是在替换这一点上，句法规则的发现注定要回到分布分析那儿去寻找工具；结果，形成了关于短语结构的分布分析的两种方法：一种是连续的短语结构的分布分析，这不过是把词的分布分析推广到

① 详见翁富良、王野翊(1998)，第4—8页。

② 详见白硕(1995)，第1—5页；翁富良、王野翊(1998)，第4—8页。

短语上去；一种是不连续的短语结构（即词结）的分布分析，这就是变换分析，藉此可以发现语义关系不变的一组实词（即词结）在多种句法环境中的分布。这在方法论上，对语言学研究无疑是有很大的启示作用的。其实，在语言学界也有人想揭示分布分析跟变换分析之间的关系。比如，袁毓林（1989，§3）指出，变换分析方法是建立在结构主义关于"焦点＋语境"的分布理论之上的。变换分析无非是通过由变换式提供的各种新语境来反映某种焦点类的分布特征，揭示其语法特性及其对句式的影响。但是，袁毓林（1989）所谓的焦点类只包括句法结构中起关键作用的词和短语，白硕（1995）则推广到词结这种不连续的句法结构。不过，白硕（1995）把词结定义为超距相关的实词多元组，这显然缺少操作上的规定性。在实际工作中人们无法决定一个句子中的哪些实词应该算作一个词结。这方面，配价语法（Valency Grammar）理论也许能帮上一点忙。根据袁毓林（1998，§3.4），价反映了动词对其他词项的支配能力，具有不同的支配能力的动词有不同的价；……价反映了动词的某种分布状况——它到底能跟多少、哪些从属成分共现；或者说，价是对动词的某种分布的集约化表示——用数字来反映动词能跟多少从属成分共现（第87页）。也许，我们可以把词结具体地规定为：动词等谓词跟其从属成分构成的实词多元组。这样，配价语法关于动词等谓词性成分的价数（能支配多少从属成分）、价质（这些从属成分的语义角色是什么）、配位方式（同一谓词的语义格不同的从属成分的同现限制关系、施事和受事等语义成分跟主语和宾语等句法成分的连接关系）等，[1]都可以为进一步充实和发展词结学说提供支持。

6 结语：并非悖论——用计算机和为计算机研究语言

最近几年，国际计算语言学界对计算语言学的定义逐步形成下

① 详见袁毓林（1998）第二、三章。

面这种共识：[①]

计算语言学是用计算机和为计算机研究语言的学科。

说计算语言学的特点是"用计算机"(by computer)来研究语言，这既有其通俗易懂的一面，又有其浅显误导的一面。其通俗性表现在：人们很容易想到计算语言学是把计算机作为工具来使用的，比如用计算机收集语料、分类整理、分布统计、提取各种数据等。这跟化学、物理学、生物学中的计算化学、计算物理学、计算生物学有点相近，它们或者运用简单的方程和算法在计算机上进行大量的重复运算，或者用计算机对实验结果进行十分精细的计算分析、反复提高以得到一种新的理论。其误导性表现在：人们只想到用计算机这种电子装置作为语言研究的工具，而忽略了用计算机科学的理论、概念和方法来研究语言这一点。我们认为这一点才是计算语言学更本质、更深刻的特点。像§5介绍的白硕(1995)用理论计算机科学的观点剖析当代语言学的方法、并进行计算模拟的做法，在一定程度上展示了这类研究的理论魅力和实用价值。在这方面，计算神经科学(computational neuroscience)为我们提供了一个光辉的典范。作为神经科学的一个新的分支，计算神经科学通过建立脑模型来阐明神经系统信息加工的计算原理，以了解人和动物的神经系统是怎样使用它的微观组件及其相互作用来表征和处理信息的。具体的做法是：把神经科学对脑结构和机能从整体、细胞和分子水平上进行的生物学研究作出数学概括、找出规律和算法，并运用现代数字计算机或人工神经网络加以模拟；其最终目标是：揭露脑的电信号和化学信号，寻求如何表达和处理神经信息、并在智能活动中发生变化的规律。这种脑模拟研究通常使用简化的脑模型。因为，即使是最成功的生物脑模型也不能揭示脑组织的全部实际功能；所以，计算神经科学需要抓住重要的原理进行简化模拟。简化模型的研究必须提供建立模型的理论框架、算法及其约束条件，而这种简化模型中的算法及其约束条件往往可以通过现代数字计算机或神经计算机来加以实

---

① 这种表述方式，笔者最早是1992年左右从黄昌宁老师那儿听来的。

现。可见，计算神经科学并不意味着大量的计算，也不意味着一定要使用现代计算机，而是要对大脑的认知过程进行表征，把其信息加工过程和信息存储过程跟计算机进行类比，从中得到新的概念和数学表达。比如，Hopfield 模型的建立并没有借助计算机进行大量的数值计算，但是这种模型有助于对大脑获取信息（即学习）和提取信息（即记忆）过程的理解；因此，这种数学模拟仍是计算神经科学的一个组成部分。同样，我们认为，计算语言学并不意味着大量的计算，也不意味着一定要使用现代计算机，而是要对大脑中的语言处理过程进行表征，把语言信息的加工、存储过程跟计算机进行类比，从中得到新的概念和数学表达，以形成便于机器处理的语法规则或语法形式体系。计算神经科学致力于寻求理解智能活动的神经基础的新概念、新算法，并在把新算法及其约束条件跟当代各类计算机进行类比中，发现设计智能化计算机、智能化机器人和智能化武器的新原理。并且，计算神经科学提出的脑模型能够对神经系统的某些行为作出可以验证的预测，从而较早地预见到生物脑研究工作的成果。因此，计算神经科学对大脑的模拟研究，不仅为信息科学的发展提供了坚实的神经科学基础，而且对神经科学和心理科学的发展也起着巨大的推进作用。[①] 我们则希望，采用理论计算机科学的观点所进行的计算语言学研究，不仅对信息科学、神经科学和心理科学起推动作用，而且对语言科学的发展起巨大的推动作用。

"为计算机"（for computer）研究语言，指为了计算机能处理自然语言而研究语言，即以计算机为应用目标来研究语言。这包括两方面的工作：(1) 对自然语言的结构和意义规律进行挖掘，提炼出便于形式化和算法化的句法、语义规则，建立合适的语法学理论模型，来更好组织语言的句法、语义规则；(2) 把语言学家对语言的句法、语义、语用诸平面上的研究成果进行数学概括，用某种形式化体系来组织和表示语言的结构和意义规则，再找出恰当的算法来描述句子的结构分析或语义解释的严格的步骤，最后根据算法用相应的计算机语言来编程实现。上面(1)所说的工作本应完全由理论语言

① 详见沈政、林庶之(1992)，第 44—49 页。

学家来承担，但是，由于理论语言学关心的方面不一定跟计算语言学一致，因而计算语言学家常常会发现：语言学中并无他们想要的句法、语义规则或语法理论模型；于是，计算语言学家只得亲自动手来寻找句法、语义规则，甚至建构更适合计算机的语法理论模型。

在为计算机研究语言这一点上，计算语言学有别于计算化学和计算神经科学。在计算化学中，并没有为计算机研究化学这种任务；在计算神经科学中，也没有为计算机研究神经的结构和功能这种任务。那么，为什么计算语言学要特别地强调为计算机研究语言这一点呢？原因可能有两点：(1) 语言学的研究对象是自然语言，语言学的研究工具(用以描写语言现象、表述语言规律、总结研究结果)也是自然语言。也就是说，自然语言既是语言研究的对象语言，也是语言研究的元语言。由于计算机无法直接理解自然语言，因而首先必须把用自然语言表述的语言规律形式化、符号化。(2) 语言是一种心智(mind)现象，是跟人的认知、心理密切相关的；为了让计算机能理解自然语言，必须以计算机为信息加工模型来考察人类语言理解的心理过程，以便在计算机上模拟实现。

可见，用计算机和为计算机研究语言并不是一种悖论，而是计算机语言学的本质特征。说来也真是饶有趣味，现代语言学的创始人索绪尔(de Saussure)告诫我们：语言学的唯一的、真正的对象是就语言和为语言而研究的语言。[①] 不到一个世纪，语言学的分支学科计算语言学的特色竟是用计算机和为计算机研究语言。语言学在本世纪的深刻变迁，从中可以略见一斑。

参考文献

白　硕 (1995)《语言学知识的计算机辅助发现》，科学出版社。
范继淹、徐志敏 (1980)《自然语言理解的理论和方法》，《国外语言学》第5期。

① 见索绪尔(1980)《普通语言学教程》，高名凯译，商务印书馆，第323页。Wade Baskin 英语译本译作 the true and unique object of linguistics is language studied in and for itself.(语言学真正和唯一的对象是就语言和为语言来研究的语言)，p. 232。必须指出的是，近年来有的索绪尔文稿研究者认为，上引的那句话并非索氏本意，而是《普通语言学教程》的编集者加上去的。我们不了解事实的底细和真相，姑且录以存疑。

冯志伟 (1992)《计算语言学对理论语言学的挑战》,《语言文字应用》第1期。

冯志伟 (1996)《自然语言的计算机处理》,上海外语教育出版社。

桂诗春、宁春岩 (1997)《语言学方法论》,外语教学与研究出版社。

黄昌宁 (1990)《语料库语言学》,《中国计算机用户》第11期。

黄 奕 (1985)《认知过程的语言》,《国外语言学》第3期。

黄建烁 (1991)《计算语言学研究综述》,《国际学术动态》第4期。

陆致极 (1990)《计算语言学导论》,上海教育出版社。

马希文 (1986)《计算机和思维科学》,见钱学森主编《关于思维科学》,人民出版社。

钱 锋 (1990)《计算语言学引论》,学林出版社。

沈 政、林庶之 (1992)《脑模拟和神经计算机》,北京大学出版社。

石纯一、黄昌宁、王家廞 (1993)《人工智能原理》,清华大学出版社。

翁富良、王野翊 (1998)《计算语言学导论》,中国社会科学出版社。

杨 抒 (1988)《自然语言的认知模型》,《计算机科学》第3期。

袁毓林 (1989)《论变换分析方法》,《汉语学习》第1期。收入袁毓林(1999)《袁毓林自选集》,广西师范大学出版社。

袁毓林 (1996)《语言的认知研究和计算分析》,删节本见《语言文字应用》第1期。全文见罗振声、袁毓林主编《计算机时代的汉语和汉字研究》,清华大学出版社。

袁毓林 (1998)《汉语动词的配价研究》,江西教育出版社。

Gazdar, G. & Mellish, C. (1987) Computational Linguistics, in J. Lyons, etc. (ed.) *New Horizons in Linguistics* 2. Penguin Books.

Grishman, Ralph (1986) *Computational Linguistics: An Introduction*. Cambridge University Press.

Halvorsen, Per-Kristian (1988) Computer Applications of Linguistic Theory, in F. J. Newmeyer (ed.) *Linguistics: The Cambridge Survey*, Vol. II, *Linguistic Theory: Extensions and Implications*. Cambridge University Press.

Winograd, Terry (1983) *Language as a Cognitive Process*. Addison-Wesley Publishing Company, Inc. 中文简介请看黄奕(1985)。

1999年11月初稿,2000年2月改定

(删节发表于《中国社会科学》2001年第4期)

基于统计的语言处理模型的有用性和局限性

本文通过介绍和评论基于统计的语言处理模型的工作原理和有关的应用实例，从语言学理论的角度来说明统计模型的有用性和局限性。首先，介绍上世纪中叶在信息论影响下的对语言的统计结构的研究，特别是乔姆斯基等对于有限状态语法不适合于刻画自然语言的论证；分析有限状态语法之类的线性语法对于语言教学的不适用性，还讨论了概率统计方法对于语素和词语界线的判定、直接成分的切分点的确定、结构核心的断定等语法分析的不适用性。然后，通过讨论N元语法模型、隐马尔科夫模型、概率上下文无关语法和基于统计的语音识别、机器翻译、词类标注、歧义消解，来展示基于统计的语言处理模型的工作原理及其可能的应用领域。接着，讨论语言结构的递归性特点和语言学知识的结构依赖性特点，指出递归嵌入使得统计规律被任意数目的嵌入词语打乱，语言学知识的结构依赖性使得统计模型赖以实现的独立性假设失效。最后，指出处理语言这种混杂系统，必须走规则与统计相结合的道路。

1 引言：统计方法和规则方法的优劣之辩

黄昌宁(2002)指出，像语音识别、词性标注等中文信息处理的急需课题，并不一定要在汉语理解的基础上推进；而是可以顺应人工智能学界在方法论上从理性主义向经验主义转变的历史潮流，在传统的基于语言学和人工智能方法的自然语言处理技术以外，大胆地启用基于语料库和统计语言模型的新方法，以满足从小规模受限语言处理走向大规模真实文本处理这种实际应用的需要。文章发表之后，在语言学界引起了强烈的反响；许多学者纷纷质疑：语言学家总结出的各种语言学规则，对于语言信息处理还有没有用处？单纯依靠概率统计的方法，能否完成中文信息处理任务？或者问得更深入一点：能否从经过标注的语料库中、通过概率统计的办法，来获得真

正的语言知识？可谓议论蜂起，莫衷一是。

其实，黄老师在文章中明确指出："尽管大规模真实文本处理是一个战略目标，不等于说小规模受限语言处理，如受限领域的机器翻译、语音对话系统、电话翻译系统和其他各种基于深层理解的自然语言分析和理论研究，就不应当搞了。目标和任务的多样化也是学术（界）繁荣昌盛的一个标志。"（第78页）只是由于黄老师在文章中主要提倡用基于统计的语言模型来研究一些紧迫的大规模真实文本处理课题，并通过词性标注方面的评测结果来说明基于统计的方法比基于规则的方法优越，因而给人一种错觉：基于规则的语言处理模型在所有方面都不如基于统计的语言处理模型。显然，这不是黄老师的文章的本意，可能也是黄老师写作那篇文章时始料未及的。

当然，本文不只是想纠正这种错觉，而是要说明：从理论上讲，语言具有递归性（recursion）的结构特点；并且，语言知识具有结构依赖的（structure-dependent）特点。这两点使得任何统计方法都难以真正挖掘出系统的语言知识，于是，基于统计的语言处理模型只能在某些非结构化的语言领域奏效。下面，我们尝试从语言学理论和有关统计方法的具体实践两个方面，作出论证。

2 语言的统计结构和有限状态语法

2.1 信息论和语言的统计结构

建立基于统计的语法模型的思想，最早源于信息论（information theory）。大家知道，第二次世界大战爆发后，由于破译密码等紧迫的军事需要，有关国家投入大量的人力进行信息编码和统计的研究。这直接促成了信息论的诞生。信息论根据信息（information）、不确定性（indeterminacy）和羡余率（redundancy）等概念，提出了有效的通信（communication）管道（channel）的测量方法。Shannon & Weaver（1949：117）在他们对信息论的开创性研究中，指出这种理论对于语言学研究可能具有的意义：

诚如我们所知道的，这种跟来源(source)相联系的信息的概念，会直接促成对语言的统计结构(statistical structure of language)的研究。拿英语来说，信息似乎对于研究语言和交际的每一方面的学者，必定都是重要的。看起来，使用涉及马尔科夫过程这种强有力的理论的观念，对于语义学研究尤其有前途；因为这种理论特别适合于处理意义的最重要、但也是最困难的的方面，即语境的影响。人们有一种模糊的感觉：信息和意义也许可以证明为是一种跟量子理论中的一对标准的共轭变量一样的东西，它们具有这样的共同参与的限制：当人们专注于其中的一个时，就得牺牲掉另一个。

对于 Shannon 和 Weaver 的提议，许多语言学家作出了热烈的响应。其中，反应最热烈的要数后布龙菲尔德学派(Post-Bloomfieldian)的领军人物 Charles C. Hockett。Hockett (1953)对信息论作了具体的介绍和评论，并指出了信息论在语言学以及其他方面的可能应用；讨论了音位化(phonemicization)和信号单位的最大平均熵(maximum average entropy per single-unit)问题(p. 81)、音位系统的统计结构和总体熵问题(p. 86)、语素-音位的转换和概率问题(p. 87)，特别讨论了怎样利用语素序列的统计特点来判断直接成分的界限(p. 88)。Hockett (1955)把信息论的成果应用到关于人类语言的马尔科夫过程模型(Markov-process model)的构造中。他运用通信理论的概念和术语，把 Bloomfield (1935)中 Jill 和 Jack 的 S→r …… s→R 语言交际模式作了重新分析；指出语音单元可以看作纯粹由信源(information source)发出的离散的信号流，其数学性质可以用 Shannon 发展的技术、通过基于对信源发出的信号流的统计数据来刻画。他用状态(states)和转移概率(transition-probabilities)组成的矩阵表来说明语句的统计结构，还引入熵来度量每一种状态的不确定性。这种以马尔科夫过程模型为基础的语法模型就是下一节将要讨论的有限状态语法(finite state grammar)。这种语法的机制(device)跟通信理论家(communication theorists)所主张的类型

十分相似，而对 Hockett (1955)来说则是完全一致。[①] 他相信，如果统计英语中所有语素和许多语素序列实际出现的相对频率，进行适当的计算；那么，整个语法结构就能用上述概率转移矩阵的方式刻画出来。当然，作为一个对语言事实有敏锐洞察力的语言学家，他认识到自然语言并不完全适合用有限状态马尔科夫过程模型来刻画；因为，(i) 有限状态马尔科夫过程中的状态数目是不随时间而改变的，但人类可以不断地学习新的语素、语言系统并不是静态的(static)；(ii) 转移概率是不随时间而改变的，这跟语言事实相反；(iii) 任何状态可能再次出现，这意味着某一语素序列的出现与否，并不改变共时的语言系统；但是，这是错误的，特别对于语言的历史研究来说。他甚至假设：通过收集状态转移概率数据、进行计算，藉此用概率转移矩阵来描写英语的结构类型；然后，把它交给工程师，由他们来制造能够理解英语的机器(p. 3—14)。这种让机器来理解自然语言的思想，在当时是极为前卫的；要知道，人工智能和认知科学的概念和名称是 1956 年才正式确立的。Hockett (1955)还用信息论上的频率、熵等概念作为度量音位的功能负荷(functional load)的数学工具，给出了一系列计算公式，其直观的含义是：一个特定对立(contrast)的功能负荷是，如果它被废除那么它将失去的熵跟没有发生变化的系统的熵之比，所有对立的功能负荷的总和是整个系统的熵。也就是说，他用如果某个对立失去后它可能对系统带来的后果，来衡量这个对立的功能负荷(p. 215—8)。

可见，用统计方法来研究自然语言的思想，在描写语言学中已经萌发。下面，为了讨论的方便，我们先简要地解释马尔科夫模型和形式语法的层级体系两个概念。

2.2　马尔科夫模型和形式语法的层级体系

马尔科夫模型描述下列这类重要的随机过程：如果一个系统有 N 个状态 $S_1, S_2, \cdots, S_N$，随着时间的推移，该系统从某一状态转移到另一状态。我们将在时间 t 的状态记为 q_t。对该系统的描述通常

---

① 参考 Newmeyer (1986)，p. 2，22，中译本第 2、27 页。

需要给出系统的当前状态（时间为 t 的状态）及其之前的所有状态，系统在时间 t 的状态处于状态 S_j 的概率取决于其在时间 1，2，…，$t-1$ 的状态，该概率为：$P(q_t=S_j|q_{t-1}=S_i, q_{t-2}=S_k, \cdots)$。如果在特定的情况下，系统在时间 t 的状态只跟其在时间 $t-1$ 的状态相关，那么该系统构成一个离散的一阶马尔科夫链（Markov chain）：

$$P(q_t=S_j|q_{t-1}=S_i, q_{t-2}=S_k, \cdots)=P(q_t=S_j|q_{t-1}=S_i) \tag{1.1}$$

进一步，如果只考虑公式（1.1）独立于时间 t 的随机过程：

$$P(q_t=S_j|q_{-1}=S_i)=\alpha_{ij}, \ 1 \leqq i,j \leqq N \tag{1.2}$$

该随机过程为马尔科夫模型。其中，状态转移概率 α_{ij} 大于或等于 0，并且 N 个状态之间的转移概率之和为 1。马尔科夫模型可以看作随机有限状态自动机（automata），其中每一个状态转换都有一个相应的概率，用以表示自动机采用这一状态转换的可能性。①

直观地说，形式语言是用来精确地描述语言及其结构的形式化手段，它以 $\alpha \rightarrow \beta$ 这种重写规则（rewriting rule）的形式来表示字符串（string）的生成。如果指定一个初始符号（initial symbol），某规则以其为左部，一组规则就可以构成一个语法。由一个语法生成的所有字符串便是语言。Chomsky（1956）根据重写规则的表达能力，区分了有限状态语法、短语结构语法（phrase structure grammar）和转换语法（transformational grammar）三种语言描写模式。在此基础上，理论计算机科学家根据对产生式（production）附加的限制条件的不同，定义了四类语法：正则语法（regular grammar）、上下文无关语法（context-free grammar）、上下文有关语法（context-sensitive grammar）和无限制重写系统（unrestricted rewriting system）；并把这四种结构表达能力不同的语法，称为“乔姆斯基层级体系”（Chomsky hierarchy）。相应地，由这些语法生成的语言是：正则语言、上下文无关语言、上下文有关语言和递归可枚举语言。严格地说，形式语法是一个四元组 $G=(N, V, P, S)$，其中 N 是非终端符号（non-terminal symbol）的有限集合，V 是终端符号（terminal symbol）的有限集合，

① 详见翁富良、王野翊（1998），第 122—124 页。

P 是一组重写规则的有限集合，而 S 是一个特定的初始符号；$P=\{\alpha\rightarrow\beta\}$。

a. 如果 P 中的规则满足如下的形式：$A\rightarrow Bx$，或 $A\rightarrow x$；其中，A，B 是非终端符号，x 是终端符号；那么称 G 为正则语法（regular grammar），即有限状态语法（简称 FSG）。

b. 如果 P 中的规则满足如下的形式：$A\rightarrow\alpha$；其中，A 是非终端符号，α 是由 N 和 V 中字符所组成的字符串（可以表示为 $\alpha\in(N\cup V)^*$，星号表示它右边的字符可以重复 0 到任意多次）；那么称 G 为上下文无关语法（简称 CFG）。

c. 如果 P 中的规则满足如下的形式：$\alpha A\beta\rightarrow\alpha\gamma\beta$；其中，$A$ 是非终端符号，α,γ,β 是字符串，且 γ 中至少包含一个字符；那么称 G 为上下文有关语法（简称 CSG）。

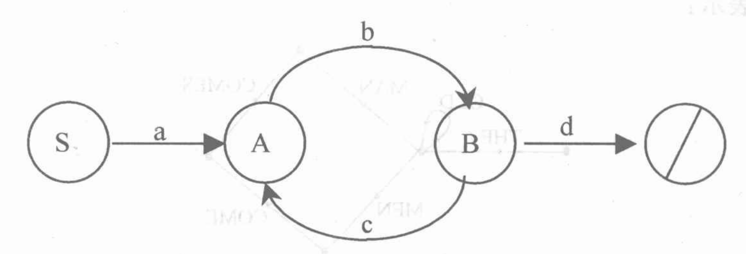
d. 如果 P 中的规则满足如下的形式：$\alpha\rightarrow\beta$；其中，α,β 是字符串，并且对产生式 $\alpha\rightarrow\beta$ 不附加任何条件；那么称 G 为无限制重写系统（简称 URS）。

给定一部语法，其相应的语言定义为所有合法终端字符串的集合。合法终端字符串指由初始符号 S 出发，运用重写规则而派生得到的终端字符串。用以规则形式表示的语法来定义语言的好处是简单明了、各成分之间的关系清楚。但是，不易判定一个字符串是否属于这套规则所规定的语言。这时，通常要借助自动机来做这种识别（recognition）工作。因为自动机可以用来机械地刻画对输入字符串的处理过程，只要根据转移函数所规定的状态和动作进行操作：如果达到终止状态，那么认为该字符串属于此语言。自动机可以根据其识别能力（源于其能够使用的信息存储空间）分为四类：有限状态自动机（finite state automata，简称 FSA）、下推自动机（push-down automata，简称 PDA）、线性界限自动机（linear bounded automata，简称 LBA）和图灵机（Turing machine，简称 TM）。它们的识别能力依次递增，分别对等于上文提到的四类语法。①

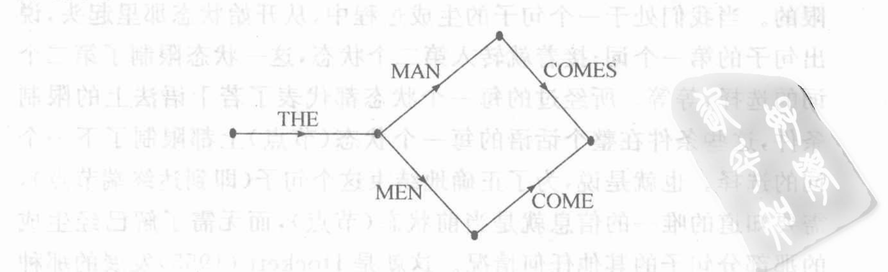
① 详见翁富良、王野翊（1998），第 34—45 页；更具体的讨论，请看张立昂（1996）。

2.3 有限状态语法和有限状态语言

具体地说，有限状态语法（正则语法）是一种线性语法（linear grammar），分为左线性语法（left-linear grammar）和右线性语法（right-linear grammar）两种。在左线性语法中，在重写规则的右侧，单独的非终端符号只能位于单独的终端符号的左侧；在右线性语法中，则正好相反。对于一部正则语法，我们总能用信息论所建议的有限状态转移图（finite state transition diagram）来表示。比如：



图上每个带标记的节点对应于一个非终端符号，最右侧的一个特殊的节点叫终端节点，用带斜线的圆表示。每一个节点对应于生成中的一个状态。为了生成被正则语法所定义的语言中的一个句子，只需在跟它对应的有限状态转移图上，从起始点开始，沿着任何一条弧从当前节点转移到下一个新节点，并记下该弧上标注的符号。当到达最后节点时，我们所记下的符号串就是这种语言的一个句子。换句话说，在状态转移图上每一条从起始点到最后节点的路径都对应于被这部语法所生成的语言中的一个句子。比如，只能生成 The man comes. 和 The men come. 这两句话的语法可以用下列状态图来表示：



我们可以给这部语法增加若干封闭圈(closed loops)加以扩展,就可以生成无限数的句子。这样,除了上列句子外,还包含下列句子:

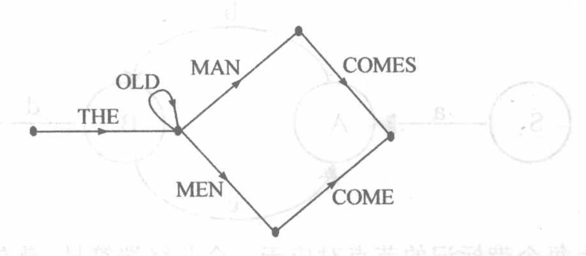
(1) a. The old man comes.

b. The old old man comes.

(2) a. The old men come.

b. The old old men come.

包含这些句子的局部的英语有限状态语法,可以用下列状态图来表示:



可见,从一个状态到另一个状态,允许有几条不同的途径;也可以随便加多少个封闭圈,并且封闭圈的长度不限。这种生成语言的机制在数学上叫做"有限状态马尔科夫过程"(finite state Markov process)。为了完成这个根据初级的通信理论编制的语言模式,我们可以给每一个状态转移加上一个概率,这样就可以计算每一个状态所带的"不确定性"(uncertainty);再用这个不确定性的平均数跟相连的各个状态的概率加权,就可以得到语言的"信息内容"(information content)。于是,通过这种概括就可以来研究语言的统计结构。显然,"有限状态"指的是状态转移图中的状态(节点)数量是有限的。当我们处于一个句子的生成过程中,从开始状态那里起头,说出句子的第一个词;接着就转入第二个状态,这一状态限制了第二个词的选择,等等。所经过的每一个状态都代表了若干语法上的限制条件,这些条件在整个话语的每一个状态(节点)上都限制了下一个词的选择。也就是说,为了正确地结束这个句子(即到达终端节点),需要知道的唯一的信息就是当前状态(节点),而无需了解已经生成的那部分句子的其他任何情况。这就是 Hockett (1955)发展的那种

模式。

作为一种形式化体系(formalism),这种语法利用有限状态网络,来为生成和分析语言提供简单的机制。但是,它不能生成许多特殊而有趣的语言。一个简单的例子是,字母"x"两边围以任意数目的成对括号:

(3) x, (x), ((x)), (((x))), ((((x)))), …

为了生成这种语言的一个句子,当生成到"x"时,我们必须知道前面已经生成了多少个左括号"(",以便能生成同样数量的右括号")"。因此,这种语言无法由正则语法来生成。[①]

2.4 自然语言不是有限状态语言

乔姆斯基在五十年代中后期的一系列研究,宣告了有限状态语法对于描写人类语言不适用。Chomsky (1956,1957)证明,没有一个有限状态语法能生成一种具有下列情况的语言:包含无限组带有嵌套依存(nested dependencies)的语符列(string),但又同时排斥无限组跟这些嵌套依存相矛盾的语符列;在英语中,也有类似上面这种嵌套结构。假如 S_1, S_2, S_3, …为英语的陈述句,那么就可以有这样一些英语句子:

(1) a. If S_1, then S_2.
 b. Either S_3, or S_4.
 c. The man who said that S_5, is arriving today.

其中,(1a)中的 then 不能用 or 来代替,(1b)中的 or 不能用 then 来代替,(1c)中的 is 不能用 are 来代替。显然,各句逗号两边的词之间,都有相互依存的关系(即 if-then, either-or, man-is)。但是,在每句相互依存的词之间,都可以嵌入一个陈述句 S_1, S_3, S_5;并且,这个陈述句实际上可以是(1a-c)中的某一句。如果把(1b)代替(1c)中的 S_5,将得到:

---

① 本节内容详见 Chomsky (1957), p. 18—25,中译本第 12—19 页;Newmeyer (1986), p. 22—26,中译本第 26—32 页;石纯一 等(1993),第 341—343 页。

(1) c'. The man who said that either S_3 or S_4, is arriving today.

把(1c')代替(1a)中的 S_1,将得到:

(1) a'. If the man who said that either S_3 or S_4 is arriving today, then S_2.

因此,一部用来生成英语句子的程序必须记住,当它经过 S_3 时前面曾经生成过什么样的构造,以便为了跟 either 匹配而生成 or、为了跟 if 匹配而生成 then。这个问题跟要生成成对的括号表达式类似。这样一类构造说明,一个有限状态转移网络(正则语法),就像不适宜用来描写括号表达式一样,也不适宜用来描写英语这样的自然语言。因为,说到底,像英语这样的自然语言不是一种有限状态语言,其中包含着有限状态语法无法处理的嵌套依存结构。比如,Chomsky and Miller (1963: 286)举了这样一个例子:

(2) Anyone_1 who feels that if_2 so-many_3 more_4 students_5 whom we_6 haven't_6 actually admitted are_5 sitting in on the course than_4 ones we have that_3 the room had to be changed, then_2 probably auditors will have to be excluded, is_1 likely to agree that the curriculum needs revision.(任何人,{他感到:如果有这么多的未经许可的〔比得到许可的学生多得多的〕学生坐在教室中听课、以至于那个教室必须更换,那么也许旁听生必须被拒之门外,}可能都会同意这门课程需要修改。)

其中,相同的下标数字表示这两个成分之间的依存关系。一个说英语的人能够产生并理解这种句子,说明根据马尔科夫过程之类的模式建立起来的语言结构理论是不能说明人类的这种语言能力的。[①]

---

① 详见 Chomsky (1957), p. 18—25,中译本第 12—19 页;Newmeyer (1986), p. 22—26,中译本第 26—32 页;石纯一 等(1993),第 341—343 页。

3 线性语法模式和语言教学

3.1 评价语法的理论标准和应用标准

Chomsky (1964b, 1965)从理论的角度,提出了语法描写的充分性平面(levels of adequacy)学说。他认为,就较低层面来说,“如果语法正确地表示了原始资料”,那就达到了“观察上的充分性(observational adequacy)”;就较高层面来说,“当语法能够对母语使用者的语言直觉(linguistic intuition)作出合理的解释,和(特别是)用这个语言的底层规律的富有意义的概括来指明观察到的资料的时候”,那就达到了“描写上的充分性(descriptive adequacy)”。至于“解释上的充分性(explanatory adequacy)”是语法理论本身要达到的一个层面,而不是通过理论中一个特定的描述来完成的。如果一个理论能够在具备良好的动因且具有人类语言官能根据的一个关于语言普遍现象的理论的基础上,从互相竞争的一组在描写上充分的语法中来选择一个,那么这个理论就被认为具有“解释上的充分性”。①

上面这些标准是从语法理论或理论语法的角度说的。但是,对于语言教学等实际的应用目标来说,可以用观察的完备性(observational adequacy)、描写的完备性(descriptive adequacy)和描写的简洁性作为衡量一个语法模式优劣的标准。具体地说,观察的完备性用以检验语法描写中所作的陈述和观察到的有关“事实”之间的一致程度;描写的完备性用以检验语法描写按其目标应予收罗的所有有关事实究竟能收罗到何种程度,对于一种企图揭示人类的语言能力的语法来说,应该把说本族语的人对这种语言的结构所“了解”的一

---

① 详见 Chomsky (1964a), p. 28; Chomsky (1964b), p. 924; Chomsky (1965) p. 24—27,中译本第 23—26 页。参考 Newmeyer (1986), p. 73,中译本第 92—93 页。这里的 adequacy,吴黄铭先生译作“妥当性”;徐烈炯先生告诉我,还是译作“充分性”好。我私下认为,像 Corder (1979)《应用语言学导论》的中文译者那样,译作“完备性”也不赖。

切都纳入描写的范围。描写的简洁性用以检验语法描写的效能。比如，如果一种描写能用数量较少的陈述或“规则”来解释同等数量的事实，或者能用同样数量的规则来解释更多的事实，它就可以被认为是更简洁或更有效能。当然，这些不同的衡量尺度在一定程度上并不互相依赖。一种语法描写中的陈述可能具有观察的完备性，但并不能令人满意地包括所有的有关事实；或者，虽然包括所有的有关事实，但表达却含糊不清、错误百出或表达方式很笨拙。理想的情形是，语法描写精致而又朴实；并且，概括出来的论断有重要的意义，可以让我们对语言是如何起作用的增进了解。如果用这些标准来衡量传统语法，那么我们也许可以说它们的观察的完备性“一般”、描写的完备性“良好”、但描写的简洁性则“差劲”。[①] 下面，就用这些实用性标准来衡量有限状态语法这种线性的语法模式。

3.2 线性语法的局限性

之所以称有限状态语法为线性语法，是因为这种描写模式把一种语言的句子看成一串语法范畴，就像项链上的珠子一样；或者看成一连串“空位”，需要用适合于每个空位的类别的词去填充。因此，这种模式有时也被称为“空位和填充”语法(slot-and-filler grammar)。这种类型的描写把句子的结构看成是线性的，在这线性的一连串范畴中，每一范畴的选择要依赖于紧接在它前面的那一个范畴。当然，这种线性序列的某个空位，在大多数情况下都有一个选择范围，只是有的范围大一点、有的小一点罢了；在少数情况下，可能没有选择的余地，即选择是唯一的。例如：

(1) He has {carefully/generally/not/given/to/a/…} ….

(2) He has {carefully/generally/not/ * given/ * to/ * a/…} taken ….

(3) It consists {of} ….

(4) It is bigger {than} ….

---

① 本节内容详见 Corder (1979), p. 177—83；中译本，第 167—74 页。

(5) I expect {to} ….

从例(1)可以发现,在〔he〕→〔has〕→〔 〕…这样的序列中,空位中有许多选择的可能性;比如,方式状语、频率状语、否定词 not、过去分词、to、冠词 a 和 the,等等。从例(2)可以发现,在〔he〕→〔has〕→〔 〕→〔taken〕…这样的序列中,由于已经知道了下一个将出现什么样的词(即前后都不独立),因而空位中选择的可能性就大大地减少了。从例(3)—(5)可以发现,在诸如 consist of 和 expect to 等固定词组、或形容词的比较级之后的连词等情形下,这种序列的空位中只有一种选择的可能性。

上文已经指出,乔姆斯基在五十年代已经证明:用这种方式来描写一种语言中的所有的句子结构在原则上是不可能的。除了这种明显的不完备之外,这种类型的描写对于理论目的和应用目的来说,还有一些其他的缺点和不合适之处。首先,它显然是极其笨拙和不简洁的、并且缺乏效能,表现在概括只能达到很低的水平。这种水平的概括只比一份清单(list)略胜一筹,是最不完备的描写。这种描写不容许具有不同程度的相似或差别,句子成分要么被看成完全相同、要么被看成完全不同。例如:

(6) 〔These〕→〔boys〕→〔paint〕…
(7) 〔This〕 →〔boy〕 →〔paints〕…

根据这种线性语法模式的描写方式,(6)(7)这两个序列必须看成是根本上有区别的。这样,它就把传统语法认为是同一范畴的语法事实(比如,名词和代词的单复数形式、同一动词的不同的人称形式)看成是彼此毫无联系、而且完全不同的范畴。这不仅是不简洁的,而且跟说本族语的人的直觉相抵触。因此,这种语法模式既缺乏效能,又缺乏描写的完备性。其次,它不能简洁地和完备地处理传统语法上所说的一致关系;具有一致关系的范畴虽然并不前后紧接在一起,但是它们之间有相互依存的关系。例如:

(8) a. 〔The〕→〔boy〕 →〔generally〕→〔paints〕→ …
b. 〔The〕→〔boys〕→〔generally〕→〔paint〕 → …

(9) 〔The〕→〔boy〕 / 〔The〕→〔boys〕 } →〔generally〕{ →〔paints〕→ … / →〔paint〕→ …

(10) a. ＊〔The〕→〔boy〕→〔generally〕→〔paint〕→ …

b. ＊〔The〕→〔boys〕→〔generally〕→〔paints〕→ …

这种线性的语法模式无法对(8)中的两个序列进行概括,使之成为(9)这类更具概括性的序列。如果那样做的话,就等于也得承认(10)中的两个序列是可容许的,即符合语法的。因为,这种语法模式中的每一个空位上可能出现的形式仅由其前面紧邻的成分决定,而副词 generally 后面出现 paint 或 paints 都是可能的。于是,频率副词可以插入到主语和谓语动词之间这样一条为每一个说本族语的人都"知道"的语法规则,就不可能用这种语法模式来描写;最终,必定导致观察上的不完备性。

不仅如此,反对这种描写模式的主要原因在于:它对句子成分之间的各种关系不加区分,都看成是相同的。因此,它无法说明一个句子的语法是怎样使我们得以解释这个句子的。任何旨在获得描写完备性的语法,必须试图将句子的结构(句子成分及其关系)跟句子的意义联系起来,而线性语法却做不到这一点。①

3.3　线性语法在语言教学上的失败

值得注意的是,这种模式曾经以一种更概括的形式,被用作编写语言教材的基础。Fries (1957)是使用这一模式的最有名的例证。在这种描写中,英语所谓的"基本"句子作为一"连串"的成分范畴("词序"模式或"公式"),被列举出来。比如,他称下面的句型为公式Ⅰ:(我们对他的标记法作了修改,以便印刷)

(1)〔DS〕+〔1+〕+〔(2+/−)−d〕+4

其中,DS=限定词,1+=复数名词,(2+/−)−d=复数或单数

① 本节内容详见 Corder (1979), p. 177—83;中译本,第 167—74 页。查核原文 p. 181,发现:中译本第 172 页上的图 12 与图 13 的次序弄颠倒了,图下的英语句子的动词后漏了词尾-s。

动词的过去式，4＝副词。根据这个公式，将生成下面这样的句子：

(2) a. The pupils ran out.

b. The ships sailed away.

这种句型在课本以“替换”表(substitution table)的形式出现，如下表所示：

限定词	名　词	动　词	副　词
The My These	pupils cows men	ran walked went	out away home

这张特定的表格，将生成34个合乎语法的序列。经过扩展后，它还可以试图把一致关系中的从属关系也包括进去。例如：

<table>
<tr><th>限定词</th><th>名　词</th><th>副　词</th><th>动　词</th><th>副　词</th></tr>
<tr><td>These
Some</td><td>boys
men</td><td rowspan="2">generally</td><td>walk</td><td rowspan="2">home
away
out</td></tr>
<tr><td>This
A</td><td>boy
man</td><td>walks</td></tr>
</table>

从教学的角度反对这种描写，其理由只是在于：除了它在描写上明显的不完备之外，也缺乏概括；因而意味着学习语法就是熟记一大串各不相同的而又互不相关的序列，再加上一大堆各不相同的而又互不相关的范畴。这样，任何描写都必须区分数量相当多的基本语法范畴，但不把诸如动词的现在和过去形式、形容词的阳性和阴性形式、名词的单数和复数以及定冠词和不定冠词，归纳在一起作出描写，这必定会给学生增添不必要的记忆负担。根据心理学等的研究，概括在学习中起相当重要的作用。因此，任何缺乏具有重要意义的概括的描写，都无助于学生发现这些概括，也不能指导学生应该怎样解释句子。①

① 本节内容详见 Corder (1979), p. 177—83；中译本，第167—74页。

4 统计概率方法在语法分析上的失利

4.1 用音素的共现概率来断定语素和词的界限

从上面的讨论可以看出，用基于概率的线性语法模式来系统地描写一种语言的结构，在原则上是很困难的。但是，这并不意味着概率统计方法在语言分析中毫无用武之地。并且，在历史上的确有不少语言学家尝试用概率统计等方法来确定语素或词的界限，乃至用以确定语法结构的层次和结构关系。

Hockett (1953)指出，在构成话语的一连串语素中，某个语素后面可能出现什么语素的不确定性的程度，在理论上是可以计算的。如果后面可以出现的语素的数目越大，那么不确定性就越大。在后面可以出现的语素的数目相同的情况下，如果这些语素的出现概率接近相同，那么不确定性就大；如果这些语素的出现概率相差很大，那么不确定性就小。于是，在对句法结构进行直接成分切分时，作为一种理论上可能是最佳的程序，语言学家可以在不确定性最大的地方切第一刀。比如，在 red hats 中，red 后面出现什么成分的不确定性大于 red hat；因为，red hat 后面只能出现-s，-ed 等极少数成分。[①]基于同样的原理，可以通过计算不同音素（或拼音符号）后面可能出现的音素的数量，来识别语素或词的界限；因为，一个语素或词内部各音素之后可能出现的音素的数量，一定大大地少于语素或词之后的(p. 87—8)。我们不妨替他举几个例子，来讨论一下这种设想的可行性如何。例如：

(1) two jars of shaving cream

---

① Hockett (1953)指出，在 hermetically sealed 中，可能 hermetic 之后的不确定性比 hermetically 之后的要大；但是，正确切分的第一刀在 hermetically 之后(p. 88)。我们认为，作为形容词 hermetic 之后的确可以出现许多成分，但是，它有另外一个交替形式 hermetical；于是，hermetic 之后出现-al 的概率是极高的，而 hermetical 之后出现-ly 构成副词的概率也是极高的。这样，hermetically 内部各语素之后的不确定性肯定小于 hermetically 之后的。

(2) When I was younger I enjoyed such things more.

(3) cranberry, strawberry, raspberry, blackberry, blueberry

(4) sister, brother, father, mother, daughter

可以设想,在 jars 中,j 和 ar 共现的概率要高于它们跟 s 的共现概率;在 younger 中,y, ou 和 ng 共现的概率要高于它们跟 er 的共现概率。在 shaving 中,sh, a 和 v 共现的概率要高于它们跟 i 和 ng 的共现概率;同样,i 和 ng 的共现概率要高于它们跟 sh, a 和 v 共现的概率。在 enjoyed 中,e, n, j, o 和 y 共现的概率要高于它们跟 i 和 d 的共现概率;同样,i 和 d 的共现概率要高于它们跟 e, n, j, o 和 y 共现的概率。问题是,像 j, ar, s, er, y, ou, ng, i, d, o 等广泛地出现在不同的语素中,这使得对它们的出现频率或共现概率的统计会得不出什么有意义的结论。并且,这种统计未必能把(3)中的 berry 识别为一个语素,但很有可能把(4)中的 er 误判为一个语素。因此,好像没见有人真的按照这种程序去确定语素或词的界限。[1]

4.2 用概率计算来决定直接成分的界限

Chatman (1955: 382) 阐述并发挥了 Hockett (1953: 87)的有关思想:通信工程师已经告诉我们,在任一符号串中,接下来将要出现什么的不确定性(indeterminacy)的程度在理论上是可以计算的。这种观念对于直接成分(immediate constituent,简称 IC)分析也许是有用的,因为在一段话语中,不确定性最大的点牵涉到最大的结构分岔。于是,在直接成分分析中,我们能否简单地说:后续环境的可能变化越大(即可以直接跟在一个语素后面的可能的语素替换类越多),把这个语素跟其后的成分在结构上分开的可能性越大。在这种思想的指导下,Chatman (1955) 根据概率论和信息论的有关原理,特别是符号串中的不确定性原理(the "indeterminacy in string" principle);具体地构造并实施了一种通过概率计算来确定直接成分

① 好像 Z. Harris 的哪篇文章中也曾简略地提及这种方法,出处失记。

的层次划分的方法，简称概率计算法。

范继淹(1964/1983)把这种方法的要领总结为：考察一段话里的每一个语素，计算它们之后可能连接多少个不同的语素类。然后，假设可能连接的语素类最多的地方(即出现语素类的不肯定性最大的地方)，就是结构上的交接点，直接成分应该从这里切开。切下来的各个部分又根据同一方法切分，直到所有的语素都切开为止。比如，Chatman (1955: 383—5)认为，对英语的句子 The boy played near the house 划分层次，先要进行如下计算：①

The 之后可能出现名词、动词、形容词以及 1 组功能词，共 4 类；

The boy 之后可能出现名词、动词、形容词、副词以及 11 组功能词，共 15 类；

The boy play 之后可能出现形容词和 2 种黏着形式，共 3 类；

The boy played 之后可能出现名词、动词、形容词、副词以及 6 组功能词，共 10 类；

The boy played near 之后可能出现名词、形容词、副词以及 4 组功能词，共 7 类；

The boy played near the 之后可能出现的语素类跟句首的 The 相同，共 4 类。

显然，The boy 之后有 15 种可能，不确定性最大；所以第一层次应该在这儿切开，其余类推。于是，全句的层次如下(数字表示不确定性的大小)：

4 15 3 10 7 4

The‖boy|play|||ed‖near|||the||||house.

---

① Chatman (1955)根据的是 Fries (1957)对英语词类的划分法，即把英语词类分为 I 类词、II 类词、III 类词和 IV 类词四个大类，以及十五组功能词。这四个大类实际上就是传统语法上的名词、动词、形容词和副词，而功能词则是传统语法上的虚词。为了便于理解，范继淹(1964/1983)把它们改为用传统语法的名称来叙述。现在，我们根据范继淹(1964/1983)第 224—225 页的有关表述。

范继淹(1964/1983)从理论上对这种概率计算法进行了分析,指出:所谓不确定性大就是能进入的环境最多,实际上是在计算整段话语中每一类语素出现的条件概率。条件概率对于预期语言的信息有一定的参考作用,但它是否能决定语素序列的层次组合还需要证明。我们还没有看到说明这两者之间具有必然关系的任何证据。其次,所谓不确定性的大小既然是按照语素类别来计算,那么根据不同的分类法就可以得出不同的结果。如果我们用传统的英语八大词类来计算,恐怕上例的层次组合就要改观。于是,语言本身所固有的层次结构居然会随着不同的语法体系对词类的不同划分而发生变化。这在理论上是讲不通的。再次,概率计算只能得出一种结果,不可能同时计算出两组不同的数字。于是,“年老的|男人和女人～年老的男人|和|女人”、“咬死了|猎人的狗～咬死了猎人的|狗”等歧义句式的层次差别就无法揭示出来。拿汉语的例子来检验一下,可以发现:它对同类成分组成的长串组合无能为力。例如:

(1) 北京大学 中文系 语言学专业 研究生

其中的各组成成分都是名词性的,它们之后可能出现的语素类的数量相等,即不确定性的大小一样,怎么划分层次呢?同样地,长串的动词组合也会碰到这种问题。即使是由不同的形式类构成的组合,照样也会碰到这种问题。以数量名组合“一朵花”为例,根据北京大学《现代汉语》(1962 年版)的词类划分计算如下:

“一”之后可能出现的语素类:〔1〕名词,如:一人;〔2〕代词,如:一这样想,就……;〔3〕数词,如:一百;〔4〕量词,如:一个;〔5〕动词,一看;〔6〕形容词,如:一大把;〔7〕副词,如:一不抽烟,二不喝酒;〔8〕介词,如:一从上海来,就……;〔9〕连词,如:一和二;〔10〕助词,如:一的平方;〔11〕语气词,如:一呀,二呀……。“一朵”之后可能出现的语素类:〔1〕名词,如:一朵花;〔2〕代词,如:一朵这样的花;〔3〕数词,如:一朵五个花瓣;〔4〕量词,如:一朵朵的红花;〔5〕动词,一朵开了,一朵谢了;〔6〕形容词,如:一朵红的;〔7〕副词,如:一朵很大的花;〔8〕介词,如:一朵从树上掉下来的花;〔9〕连词,如:一朵和二朵;

〔10〕助词，如：一朵的花瓣是红的；〔11〕语气词，如：一朵呀，两朵呀……。

“一”和“一朵”之后可能出现的语素类都是11类，那么应该在哪里划分呢？还有一种情形，黏着形式总起来算一个，还是有一个算一个？如果作为一类，那么“我看书”的“我”和“我看”之后都只能加上一类黏着形式。如果有一个算一个，那么“我”之后只能加一个“们”，而“我看”之后可以加“了、着、过”三种黏着形式。于是，概率计算的结果就大不一样。据此，范先生得出结论：以上的检验证实了我们的怀疑——条件概率跟层次组合并无关系，概率计算法只是一种主观的设想（第224—227页）。

4.3 用数量统计来决定结构核心

更有甚者，Pittman（1948）整理出十条标准来判定在由两个直接成分构成的序列AB中，到底哪一个是结构核心。其中，有四条用到数量统计，它们分别是：

> 标准2：类的大小。如果两个直接成分，其中有一个成分所属的形式类比另一个成分所属的形式类大（即有较多成员的类），一般把它看成是中心的（central），而把它的伴随成分看成是旁侧的（lateral）。如英语的副词与动词、代词与动词、词缀与词干。
>
> 标准3：搭配力（范围）。如果两个直接成分，其中有一个成分比另一个成分有更多不同类的伴随关系的可能出现范围，一般把它看成是中心的，而把它的伴随成分看成是旁侧的。如英语的come down，inside；法语的deux ans（两年）。
>
> 标准5：类的频率。如果两个直接成分的类，其中一个类比另一个类出现的次数多，那么就把它看成是中心的，而把它的伴随成分看成是旁侧的。如英语的名词比形容词出现的次数多、动词比副词出现的次数多、词干比词缀出现的次数多、独立句比从属句出现的次数多。
>
> 标准6：个体的频率。如果某一个个别成分出现的次数，比

它的伴随成分出现的次数多；那么就把它看成是旁侧的，而把它的伴随成分看成是中心的。如像在Nahuatl语这样的语言中，由于词干一定要带上词缀才能出现，因而某些词缀一定比词干类的任何成员出现的次数要多得多。比如，前缀ni-和后缀-tl，就比随它们一起出现的任何一个词干的频率要高得多。

显然，如果拿汉语的事实来检验，那么这几条标准都站不住。就类的大小来说，形容词的成员比动词少，在"形＋动"组合(偏正结构，如：认真学习)和"动＋形"组合(述补结构，如：码放整齐)中，正好动词是核心。形容词的成员比名词少，在"形＋名"组合(偏正结构，如：干净衣服)中碰巧名词是核心；但是在"名＋形"组合(主谓结构，如：衣服干净)中，形容词却是核心了。动词的成员比名词少，在"动＋名"组合(述宾结构，如：买汽车)和"名＋动"组合(主谓结构，如：客人走了)中，动词都是核心。就搭配能力来说，形容词比名词和动词强。比如，它可以作名词和动词性成分的谓语(如：南方湿润、去好)，这时是核心；它可以修饰名词和动词(如：小房间、努力工作)、可以作动词的补语(如：弄明白)、可以作主语和宾语(诚实好、喜欢安静)，这些情况下都不是核心。就类的频率来说，名词肯定高于动词和形容词；但是，在"名＋动"、"动＋名"和"名＋形"组合中，名词都不是核心；只有在"形＋名"组合中，名词才是核心。就个体的频率来说，一些常用名词肯定高于经常跟它们搭配的常用动词和形容词；但是，这些"名＋动"、"动＋名"和"名＋形"组合，名词也不是核心；只有在"形＋名"组合中名词才是核心。因此，Pittman (1948)也只得承认，这个标准是有例外的。……这些标准的适用程度是不同的，适用程度的大小取决于有关的语言的性质和语言学家掌握这些标准的程度。尽管如此，他还是希望它们也可以有效地用于音素结构和音节结构。而我们认为，这相当困难，无论在语法结构上、还是在音素结构和音节结构上。后来，似乎再也没有人尝试这样做，就是一个显而易见的证明。

5 基于统计的语言处理模型的工作原理

5.1 基于规则的模型和基于统计的模型

基于统计的语言处理模型是相对于基于规则的语言处理模型而言的，两者区别在于：前者是一种概率性的非确定性的语言处理模型，后者是一种确定性的语言处理模型。一般地说，确定性的模型运用明确的规则来表述物理世界（或自然语言）的已知的特定属性。在物理学中，如牛顿力学；在自然语言处理中，如正则语法、上下文无关语法等形式语法。它们都属于确定性的模型。但是，并不是所有的物理世界和自然语言的现象都可以用确定性的规则来刻划，而且这些规则的使用也具有不确定性。在这种情况下，统计模型被用以描述物理世界和自然语言的统计属性。建立统计模型的基本假设是：物理世界和自然语言可以用随机过程来刻划，而随机过程中的参数可以精确地估计。比如，物理学上的统计力学、自然语言处理中的概率语法，都属于统计模型。① 对于自然语言处理来说，它涉及的知识是海量的，尤其是当我们要面对大规模真实文本处理时，就能发现基于规则的方法会碰到下列目前还难以克服的困难：②

(1) 获取语言学知识(linguistic knowledge)以及相关的世界知识(world knowledge)是一件非常困难的事情，要想对它们进行形式化表示更不容易；更何况并不是所有的自然语言都像英语那样得到了比较深入的研究，从而具有比较成熟的句法学和语义学的描写体系。

(2) 自然语言中有大量的非单调(non-monotonous)现象，我们不能保证关于自然语言的大量的不同的规则之间一定是相容的；这样，在自然语言处理系统中，随着规则数量的增加，规则与规则之间常常发生矛盾和冲突。

---

① 详见翁富良、王野翊(1998)，第116页。

② 详见黄昌宁(2002)，第79页；白栓虎(1992)，第39页。

(3) 规则所能刻画的知识的颗粒度太大，无法用有限的规则来刻画自然语言中复杂多变的现象。

(4) 很难处理自然语言中的不确定性，比如，句子的合语法性和可接受性常常是模糊的，词和短语等语法单位的边界(boundaries)也是不十分清晰的。

因此，在目前的语言学理论水平和计算技术条件下，人们自然地会转向统计学方法，希望用在语料库中对相关数据的统计的方法，来为要解决的语言问题建立统计模型，并且由语料库中的训练数据来估计统计模型中的有关参数。于是，基于大规模语料库的概率统计模型成了自然语言处理的一种必然的选择。下面，我们先介绍三种基于统计的语言处理模型的工作原理，再介绍四个统计模型的应用实例。

5.2　N 元模型(N-1 阶马尔科夫模型)

语言的统计模型可用于计算语句 $W = w_1, w_2, \cdots, w_n$ 的先验概率 $P(W)$，在这里用变量 W 代表一个文本中顺序排列的 n 个词。根据概率论的定理(乘法规则[①])，$P(W)$可以分解为：

$$P(W) = \Pi_{i=1\cdots n} P(w_i \mid w_1, \cdots, w_{i-1}) \qquad (5.2.1)$$

其中，符号 $\Pi_{i=1\cdots n} P(\cdots)$表示概率的连乘；如果把它展开来就是：

$$P(W) = P(w_1)P(w_2 \mid w_1)P(w_3 \mid w_1, w_2)\cdots P(w_n \mid w_1, w_2, \cdots, w_{n-1}) \qquad (5.2.2)$$

可见，为了预测 w_n 的出现概率，必须知道它前面所有词的出现概率；这 $w_1, \cdots w_2, w_{n-1}$ 被称为产生 w_n 的历史。随着历史的长度的增加，不同历史数按指数级增长。如果历史长度为 $i-1$，则有 L^{i-1} 不同的历史(L 为词汇集的大小)。我们必须考虑在所有的 L^{i-1} 种历史的情况下，产生第 i 个词的概率。也就是说，这样的模型中有 L^i 个自由参数 $P(w_i \mid w_1, \cdots, w_{i-1})$。当 $L=5000$，$i=3$ 时，自由参数

---

① 乘法规则用以计算两件事一起发生的概率：两件事一起发生的几率等于第一件事发生的几率乘以已知第一件事发生的情况下第二件事发生的几率。详见 Freedman 等(1991)的中译本，第 253—255 页。

的数目是1250亿。我们几乎不可能从训练数据中正确地估计这些参数，并且绝大多数的历史在训练数据中根本没有出现。解决这个问题的方法是，将历史 $w_1, \cdots, w_{i-1}$ 按照某个法则影射到等价类 $S(w_1, \cdots, w_{i-1})$，而等价类的数目远远小于不同历史的数目。有很多方法可以将历史划分成等价类，比如，把参数空间中一些特征相近的元素合并到一起得到一个个等价类；于是，参加运算的是这些类，而不再是单个的元素。从计算上看，这样还是太复杂。如果任意一个词出现的概率只跟它前面的 $N-1(N\geqslant 1)$ 个词相关，那么问题就可以得到进一步的简化。这时的语言模型叫N元模型或N元语法(N-gram)，即

$$P(W)=P(w_1)P(w_2|w_1)P(w_3|w_1, w_2)\cdots P(w_i|w_{i-N+1}, \cdots, w_{n-1}) \approx \Pi_{i=1\cdots n} P(w_i|w_{i-N+1}, \cdots, w_{i-1}) \quad (5.2.3)$$

通常 N 的值不能太大，否则会有太多的等价类，前面提及的过多自由参数的问题仍然存在。当 $N=1$ 时，即近似地认为出现在第 i 位上的词 w_i 独立于历史（它的出现概率跟它前面的词无关）时，这种 N 元语言模型称为一元语法(uni-gram，或 mono-gram)。当 $N=2$ 时，即近似地认为出现在第 i 位上的词 w_i 的出现概率只跟它前面紧邻的一个词相关时，这种 N 元语言模型称为二元语法(bi-gram)。当 $N=3$ 时，即近似地认为出现在第 i 位上的词 w_i 的出现概率只跟它前面紧邻的两个词相关时，这种 N 元语言模型称为三元语法(tri-gram)。其实，N 元模型就是 N-1 阶马尔科夫模型。因此，一元语法就是零阶马尔科夫链，二元语法就是一阶马尔科夫链，三元语法就是二阶马尔科夫链……。当使用三元语法模型时，$P(W)$可以分解为：

$$P(W) \approx \Pi_{i=1\cdots n} P(w_i|w_{i-2}, w_{i-1}) \quad (5.2.4)$$

该模型的参数为 $P(w_3|w_2, w_1)$，其值可以通过大规模语料库、用最大似然估计(maximum likelihood estimation)方法来求得：

$$P(w_3|w_2, w_1)=f(w_3|w_2, w_1) = \text{count}(w_1, w_2, w_3)/\text{count}(w_1, w_2) \quad (5.2.5)$$

其中，count (w_1, w_2, w_3)表示一个特定的词序列 w_1, w_2, w_3 在语

料库(或训练例)中出现的次数,count (w_1, w_2)表示一个特定的词序列 w_1, w_2 在语料库(或训练例)中出现的次数,f ($w_3 | w_2$, w_1)表示在给定 w_1, w_2 的条件下出现 w_3 的概率。但是,在训练数据中,很可能事件 w_1, w_2, w_3 这种词序列根本没有出现过,根据最大似然估计,这些事件的概率为零。然而,这些事件的真实概率不一定为零。这就是所谓的数据稀疏问题(sparse data problem)。现在已经发展出解决这一问题的有关方法,此处从略。[①]

5.3 隐马尔科夫模型(Hidden Markov Model, 简称 HMM)

隐马尔科夫模型是由转移链连接的多个状态的集合。其中,每个转移链上都有两组概率:转移概率(transition probability)和输出概率密度函数(output probability density function)。前者给出了执行该转移的概率,后者定义了在执行某个转移的条件下从有限字母表中输出每个符号的概率。比较起来,在马尔科夫模型中,每一个状态代表一个可观察事件;马尔科夫模型描述的是一个随机过程(stochastic process),即状态之间的转移。这限制了模型的适用性。隐马尔科夫模型是马尔科夫模型的扩展。在隐马尔科夫模型中,观察到的事件是状态的随机函数。因此,这种模型是一种双重随机过程:一个随机过程描述输出符号与状态之间的概率关系,即输出符号是状态概率的函数;另一个随机过程描述的才是状态之间的转移关系。其中,对于外界的观察者来说,只能看见输出符号,而不能看见状态之间的转移;即该模型的状态转换过程是不可观察的(隐蔽的)。可观察事件的随机过程是隐蔽的状态的转换过程的随机函数。这种模型有如下的组成部分:(1) 模型中的状态数 N。(2) 从每一状态可能输出的不同符号数 M。(3) 状态转移概率矩阵 $A = \alpha_{ij}$,其中

$$\alpha_{ij} = P(q_t = S_j | q_{t-1} = S_i), 1 \leqq i, j \leqq N$$

① 详见翁富良、王野翊(1998),第 116—24 页;黄昌宁(2002),第 80 页;白栓虎(1992),第 50—52 页;黄昌宁、李涓子(2002),第 115—6 页。必须指出的是,翁富良、王野翊(1998),第 118 页说:“当 $N=1$ 时,……N-gram 语言模型被称为一阶马尔科夫链。当 $N=2$ 时,N-gram 语言模型被称为二阶马尔科夫链。当 $N=3$ 时,N-gram 语言模型被称为三阶马尔科夫链”。这跟一般的理解有所不同。

其中,状态转移概率 a_{ij} 大于或等于0,并且 N 个状态之间的转移概率之和为1。(4) 从状态 S_j 观察到符号 ν_k 的概率分布矩阵为 $B=b_j(k)$。(5) 初始状态概率分布 $\pi=\pi_i$。为了方便,可以把隐马尔科夫模型记为 $\lambda=(A, B, \pi)$,用以指出模型的参数集合。这种模型会碰到如何快速地计算或调节有关概率或参数、选择最优的状态序列等问题。现在已经发展出一些算法来解决这些问题,这里从略。①

5.4 概率上下文无关语法(Probabilistic Context Free Grammar,简称 PCFG)

概率上下文无关语法是上下文无关语法的概率拓广,表现为:上下文无关语法中的每一个产生式 $A\rightarrow\alpha$ 都被附加了一个概率值。对所有的非终端符号 A,该概率分布必须满足

$$\Sigma_\alpha P(A\rightarrow\alpha)=1$$

跟上下文无关语法相比,概率上下文无关语法有下列优点:(i) 在一个有歧义的概率上下文无关语法中,如果参数选择适当,正确的语法结构具有较高的概率。因而 PCFG 能够用以化解歧义(disambiguation,或 ambiguity resolution),即在诸多的语法结构(歧义分析)中选择正确的语法结构。(ii) 由于可以尽早删除语法分析过程中发现的概率很小的子结构(sub-structure),因而 PCFG 加速了语法分析。(iii) PCFG 使我们能够定量地比较两个语法的性能。给定两个语法 G_1 和 G_2,我们可以使用语料库 C 来定量地评价 G_1 和 G_2;如果 $P_{G1}(C)>P_{G2}(C)$,那么我们可以得出 G_1 优于 G_2 的结论。当然,跟隐马尔科夫模型一样,PCFG 也会碰到如何快速地计算或调节有关概率、选择最佳的语法分析等问题。现在已经发展出一些算法来解决这些问题,这里从略。② 关于这种语法存在的问题,§6.3将有简略的讨论。

5.5 基于统计的语音识别

从统计的角度看,语音识别的任务是:在给定语音讯号(speech

① 详见黄昌宁、李涓子(2002),第116—118页;翁富良、王野翊(1998),第124—136页。

② 详见翁富良、王野翊(1998),第136—44页。

signal)A 时，找出语句（即词语序列）S，使得 $P(S|A)$ 最大；也就是说，S 最可能是 A 所传达的语句。因此，语音识别可视为计算以下条件概率的极大值问题：

$$S = argmax_S P(S|A) \tag{5.5.1}$$

$P(S|A)$ 表示已知输入语音讯号序列 A 的情况下，出现词语序列 S 的条件概率。数学符号 $argmax_S$ 表示通过考察不同的候选词语序列 S，来寻找使条件概率 $P(S|A)$ 取最大值的那个词语序列 S，这后者就是当前输入语音讯号序列 A 所对应的输出词串。根据贝叶斯定律（Beys law），上式可以转写为：

$$S = argmax_S P(A|S)P(S)/P(A) \tag{5.5.2}$$

由于公式中的分母 $P(A)$ 在语音讯号 A 给定时，是一个归一化（normalization）常数，不影响极大值 $argmax_S$ 的计算；因而可以忽略不计，即把它从公式中删除。于是，得到如下语音识别的统计模型：

$$S = argmax_S P(A|S)P(S) \tag{5.5.3}$$

其中，$P(A|S)$ 叫做声学模型（或统计语音模型），它给出了从语句 S 产生出语音讯号 A 的概率。一般来说，统计语音模型是用隐马尔科夫模型来建立的。$P(S)$ 叫做统计语言模型，它给出了在一个语言中产生语句 S 的概率。

因此，语音识别系统一般由统计语音模型和统计语言模型两个部分组成。直观地说，语音识别就是搜索 S，使得这两个模型的概率的乘积 $P(A|S) \times P(S)$ 最大。[①]

5.6 基于统计的机器翻译

基于同样的原理，统计学机器翻译使用统计模型来刻画语言翻译的过程，并且自动地从平行的双语语料库中训练出这些模型的参数。其中，一个对齐（alignment）程序从双语语料库中识别出对应的句子。对于每一对相应的句子，统计学翻译系统认为它们是一个通讯信道（channel）两端的信息。如果要从法语翻译到英语，英语就是这个信道源端的发送信息，法语就是这个信道末端的接受信息。机

---

① 详见翁富良、王野翊（1998），第 20—21 页；黄昌宁（2002），第 80—81 页。

器翻译的任务变成了根据接收信息来解码,从而得出发送信息。这种思路在一定程度上正好跟传统的机器翻译相反。在这里,翻译器的任务是:在给定了法语句子 F 的情况下,搜索其相应的英语句子 E,使得后验概率 $P(E|F)$ 达到极大值。这个极值点 E^* 就是 F 的翻译。即

$$E^* = argmax_C P(E|F) \quad (5.6.1)$$

根据贝叶斯定律,上式可以转写为:

$$E^* = argmax_C P(E)P(F|E)/P(F) \quad (5.6.2)$$

由于在给定的情况下,$P(F)$ 是一个归一化常数,不影响极值点的计算。因此,任务(5.6.2)成为搜索

$$E^* = argmax_C P(E)P(F|E) \quad (5.6.3)$$

为此,一个机器翻译系统必须建立统计模型,用于刻画概率参数 $P(E)$ 和 $P(F|E)$。其中,刻画 $P(E)$ 的模型叫语言模型,通常用三元语法作为模型。刻画 $P(F|E)$ 的模型叫翻译模型,包括两种语言的词语和句子的对齐概率、在某种对齐下的翻译概率等参数。[①]

5.7 基于统计的词类标注

词类标注(part-of-speech tagging)问题,可以看作是在给定词的序列 $W = w_1, w_2, \cdots, w_n$ 的条件下,搜索词类标记序列 $C = c_1, c_2, \cdots, c_n$,使得 $P(C|W)$ 最大。即计算如下条件概率极大值的问题:

$$C = argmax_C P(C|W) \quad (5.7.1)$$

$P(C|W)$ 表示已知输入词序列 W 的情况下,出现词类标记序列 C 的条件概率。数学符号 $argmax_C$ 表示通过考察不同的候选词类标记序列 C,来寻找使条件概率 $P(C|W)$ 取最大值的那个词类标记序列 C,这后者就是对词序列 W 的词类标注的结果。根据贝叶斯定律,上式可以转写为:

$$C = argmax_C P(W|C)P(C)/P(W) \quad (5.7.2)$$

由于公式中的分母 $P(W)$ 在词序列 W 给定时,是一个归一化常数,

① 详见翁富良、王野翊(1998),第166—70页。

不影响极大值 $argmax_C$ 的计算；因而可以忽略不计，即把它从公式中删除。于是，得到下面的公式：

$$C = argmax_C P(W|C)P(C) \tag{5.7.3}$$

接着，对公式进行近似运算。首先，引入独立性假设，[①]认为词序列中任意一个词 w_i 的出现概率近似只跟当前词的词性标记 c_i 有关，而跟上下文的词类标记无关。即词汇概率（某个词以某种词类出现的概率）为：

$$P(W|C) \approx \Pi_{i=1\cdots n} P(w_i|c_i) \tag{5.7.4}$$

显然，这是一种一元语法模型，它只考虑词跟在其上可能出现的词类（标记）之间的统计信息，即一个词用作某种词类的概率。其次，采用二元假设，认为任意词类标记 c_i 的出现概率只跟它紧邻的前一个词类标记 c_{i-1} 相关。即

$$P(C) \approx \Pi_{i=1\cdots n} P(c_i|c_{i-1}) \tag{5.7.5}$$

$P(c_i|c_{i-1})$ 是词类标记的转移概率，显然这是一种二元语法模型；它只考虑词类一级上的相邻上下文关系（即某种词类序列是否出现的统计关系），但是没有考虑特定的词跟某种词类标记之间的统计关系（即一个词用作某种词类的概率）。把（5.7.4）和（5.7.5）两式代入（5.7.3），得到下面的公式：

$$C \approx argmax_C \Pi_{i=1\cdots n} P(w_i|c_i)P(c_i|c_{i-1}) \tag{5.7.6}$$

这个公式可以看作是一个隐马尔科夫模型，模型中的每一个状态对应于一个词类标记；从状态 S_i（对应于词类标记 c_i）到状态 S_j（对应于词类标记 c_j）的转移概率（a_{ij} 为相应的词类标记的二元语法模型 $P(c_j|c_i)$；从状态 S_i 输出词 w_i 的输出概率 $b_i(w_i)$ 为基于一元语法的词汇概率 $P(w_i|c_i)$。于是，词类标注问题变为求隐马尔科夫模型的最佳状态序列的问题。这种问题可以用韦特比算法（Viterbi algorithm）来解决。（5.7.6）这个公式（即隐马尔科夫模型）中的两个概率参数都可以通过训练数据（即带词类标记的语料库）来分别估计：

---

① 如果给定第一件事，无论它的结果是什么，第二件事的机会都一样；那么，这两件事是独立的。否则，就是不独立的。如果两件事是独立的，那么这两个事件都发生的机会等于它们各自无条件概率的乘积。详见 Freedman 等（1991）的中译本，第 256—258 页。

$$P(w_i \mid c_i) \approx count(w_i, c_i) / count(c_i) \tag{5.7.7}$$

$$P(c_i \mid c_{i-1}) \approx count(c_{i-1} c_i) / count(c_{i-1}) \tag{5.7.8}$$

公式(5.7.7)说的是，词汇概率约等于：训练数据中某词 w_i 作某种词类 c_i 使用的次数，除以该词类标记 c_i 在训练数据中出现的次数。公式(5.7.8)说的是，转移概率约等于：训练数据中某种词类标记 c_i 出现在另一种词类标记 c_{i-1} 之后的次数，除以另一种词类标记 c_{i-1} 在训练数据中出现的次数。①

因此，词类标注系统一般由词汇概率模型和转移概率模型两个部分组成。直观地说，词类标注就是搜索词类标记序列 C，使得这两个模型的概率的乘积 $P(w_i \mid c_i) \times P(c_i \mid c_{i-1})$ 最大。据 Garside 等(1989)报导，他们用上述方法自动标注英语词类的正确率达到96%。据白栓虎(1992：61)报导，他尝试用不同的模型来自动标注汉语词类：单纯用词汇概率(一元语法)模型时正确率达到88.3%，单纯用转移概率(二元语法)模型时正确率达到89.5%，用这两种概率的乘积(隐马尔科夫模型)时正确率达到95.2%。

5.8　基于统计的歧义消解

一般地说，歧义(ambiguity)指的是：一个句法结构可以作不同的语法分析，从而有不同的语义理解。例如：

(1) a. The boy saw the girl with a telescope.

b. The boy saw the girl with a telescope on the hill.

(2) Perre Vinken, 61 years old, joined the board as a nonexecutive director.

(1a)中的介词结构 with a telescope，既可以附加在动词 saw 上作状语，又可以附加在名词 girl 上作定语。(1b)中的介词结构 on the hill，既可以附加在动词 saw 上作状语，又可以附加在名词 girl 上作定语，甚至可以附加在名词 telescope 上作定语；当介词结构 with a

① 详见翁富良、王野翊(1998)，第170—174页；黄昌宁(2002)，第81—2页；白栓虎(1992)，第49—53、61—62页。

telescope 和 on the hill 的不同分析综合在一起时，(1b)就有多种语义解释。从结构(structure)上看，(2)跟(1a)一样是有歧义的；从词汇插入后的实例(instance)来看，人们根据词汇之间的语义限制，只从(2)上得到一种语义解释；但是，对于机器而言，(2)仍是有歧义的，可以给出两种结构描述。由于介词结构附加(PP attachment)造成的歧义具有普遍性和典型性，因而大多数统计学歧义消解的研究集中在这一问题上。

为了用计算机来求解介词结构的正确附加，必须用一种合适的知识表示(即形式化)来描述这个问题。比如，首先，用词类范畴来表示输入句子的骨架，于是介词结构的附加问题可以简化为：在具有 $verb\ np_1(prep\ np_2)$ 形态的语句中，介词结构($prep\ np_2$)应该附加于 $verb$(动词)还是 np_1(名词短语)的问题。接着，用随机变量 A 表示 $prep\ np_2$ 的附加，A 的值可取 VB(指附加于动词)或 NP(指附加于名词)；最后，用 w 表示语句中除了 $verb\ np_1(prep\ np_2)$ 之外的词。在这种情况下，如果有一个已经标注了句法结构的语料库作为训练数据(比如，宾州大学树库，Upenn Treebank)可资利用，那么就可以由概率分布 $P(A|prep, verb, np_1, np_2, w)$ 来确定某一种附加的可能性。引入独立性假设，假定上述概率分布独立于 w，并独立于 np_1，np_2 中除中心名词(head noun)以外的其他部分；那么

$$P(A|prep, verb, np_1, np_2, w) \approx P(A|prep, verb, noun_1, noun_2) \quad (5.8.1)$$

其中，$noun_1$ 是 np_1 的中心名词，$noun_2$ 是 np_2 的中心名词。如果：

$$P(NP|prep, verb, noun_1, noun_2) > P(VB|prep, verb, noun_1, noun_2)$$

那么，判定介词结构附加于名词短语；否则，判定介词结构附加于动词。

由于用上述模型来刻画介词结构的附加时，是用个别词而不是词类来进行统计的，因而该模型中有太多的自由参数。比如，假定日常英语中有 10^4 个名词、10^3 个动词、10 个介词，那么该模型大约有 10^{12} 个自由参数。事实上许多事件(即某些名词或动词跟某些介词结构的搭配序列)在语料库中根本没有出现过，这会导致严重的数据

稀疏(即训练不足)问题。所以在实际的实现中,可以采用后撤(back-off)算法。当找不到四个中心词的四元组时,就退一步找一个三个中心词的三元组,以此类推;直至退到一元组时,只根据具体的介词来作出判断。这是一种用低元语法来平滑高元语法,从而解决数据稀疏问题的方法。或者,直接对模型 $P(A|prep, verb, noun_1, noun_2)$ 作进一步的简化,作出如下独立性假设:

$$P(A=NP|prep, verb, noun_1, noun_2) \approx P(NP|prep, noun_1)$$

$$P(A=VB|prep, verb, noun_1, noun_2) \approx P(VB|prep, verb) \qquad (5.8.2)$$

也就是说,假定介词结构的附加跟介词后面所跟的名词不相关;并且,介词短语附加于宾语时跟动词不相关,介词短语附加于动词时跟宾语不相关。还可以进一步假设:

$$P(A=NP \mid prep, noun_1) > P(A=VB \mid prep, verb)$$
$$\Updownarrow$$
$$P(prep \mid noun_1) > P(prep \mid verb) \qquad (5.8.3)$$

也就是说,简单地通过比较介词跟中心语名词的共现概率与介词跟动词的共现概率的大小,来估计哪一种附加的可能性更大。根据最大似然估计原理,

$$P(prep|noun_1) \approx count(prep\ Attach\text{-}to\ noun_1)/count(noun_1) \qquad (5.8.4)$$

$$P(prep|verb) \approx count(prep\ Attach\ to\ verb)/count(verb) \qquad (5.8.5)$$

像 $count(prep\ Attach\ to\ noun_1)$ 和 $count(prep\ Attach\ to\ verb)$ 等概率参数,可以在标注了语法结构关系的树库中得到。也可以通过非歧义的数据来估计,比如,当一个名词 n 前面没有动词,后面跟有一个介词结构 PP(介词为 p)时,将 $count(prep\ Attach\text{-}to\ noun_1)$ 加一;当一个介词结构 PP(介词为 p)前面既有动词 v,又有名词短语 np,但该 np 是一个代词时,将 $count(prep\ Attach\text{-}to\ verb)$ 加一。通

过诸如此类的规则来获得相关的概率参数。[①]

6　语言的递归性和语言学规则的结构依赖性

6.1　语言的递归性和语言官能

众所周知，语言在结构方式上具有递归性（recursion）的特点，突出地表现为：一个按照某种语法模式造成的语法组合，其直接成分可以也是按照这种结构模式（或其他结构模式）造成的语法组合。[②] 比如，§2.4 中的例（1）（2）诸例；再如：

（1）The mouse the cat the dog chased bit died.

（2）a. The man who …… is here.

b. I saw a …… house.

（3）这件事儿，我们几个人中间，小王现在态度最不明朗。

（4）我不知道小李知道不知道她丈夫已经知道她没有通过律师资格考试。

当然，像例（1）这样极端的例子在真实的语言交际中是不常见的；但是，我们不能保证一定碰不到像例（3）（4）那样的句子。正如 Chomsky (1957: 17)所说的，像例（2）的……处可以分别插入任意长度的动词性词组和形容词性词组。如果用产生式规则（production rule）来表示短语结构的形成过程，那么递归性就表现为箭头左侧的符号可以出现在箭头的右侧，甚至连初始符号 S 也可以出现在箭头的右侧。[③] 例如：

（5）i. S → NP＋VP

---

① 详见翁富良、王野翊（1998），第 177—180 页；黄昌宁（2002），第 82—83 页。

② 参考 Hockett (1958)，中译本第 194—200 页。

③ 参考 Hartmann, R. R. K. & F. C. Stork (1972) *Dictionary of Language and Linguistics* (Applied Science Publishers Ltd. London)中的 recursiveness（循环性）条目。见中译本《语言与语言学词典》（黄长著、林书武、卫志强、周绍珩译，李振麟、俞琼校，上海辞书出版社，1981 年），第 292 页。另外，徐烈炯先生在岳麓论坛（2002）上也提及这一点。

ii. NP → Det＋N

iii. NP → Det＋N＋(S)

iv. VP → V＋NP

v. VP → V＋to-VP

在(5iii)中，初始符号 S 可以作名词 N 的定语从句；这样形成的名词性成分可以分别代入(5i)中的 NP 和(5iv)中的 NP，再把(5iv)这样的动词性成分代入(5i)中的 VP，就构成下面这种句子：

(6) The man (who kicked the ball) scored the goal (that won the game).

(踢球的那个人踢进了赢得这场比赛的一个球。)

递归性是人类语言的一个非常重要的特性，它把人类语言跟其他动物的交际符号系统区别开来。Hauser，Chomsky and Fitch (2002)甚至认为递归性是反映人类语言官能(the faculty of language，简称 FL)的基本属性。他们把语言官能分成广义和狭义两种，广义的语言官能(FL in broad sense，简称 FLB)包括内在的运算系统(computational system)和两个内部的有机体系统："感觉-运动"(sensory-motor)系统和"概念-意向"(conceptual-intentional)系统。狭义的语言官能(FL in narrow sense，简称 FLN)只包括抽象的语言运算系统，即狭义句法(narrow syntax)。这些不同的系统的关联方式是：运算系统生成内部表达，并通过音系系统把它们映射到感觉-运动接口，通过(形式)语义系统把它们映射到概念-意向接口。FLN 的基本属性是递归，它用有限的元素集合来造成潜在的无限的离散的表达。每一个这种离散的表达被送到感觉-运动系统和概念-意向系统，再由它们在使用语言的过程中来处理和细化这些信息。虽然 FLN 具有递归能力，但是 FLN 或 FLB 之外的许多内部机体因素对使用这个系统施加实际的限制。比如，肺容量限制了实际口语句子的长度，工作记忆对句子的复杂性施加限制以利于句子可以被理解。从现有的研究来看，FLB 的许多方面是人类和其他脊椎动物共有的，但是属于 FLN 的核心的递归方面在动物交际和可能的其他领域中都没有任何类似物(p.1570—1571)。由于有限状态语法只能

反映局部的依存关系(local dependencies),因而它不能充分地抓住任何人类语言。因为自然语言可以通过在短语中递归地嵌入短语来超越纯粹的局部结构,从而导致统计规律(statistical regularities)被任意数目的词或短语打乱。这种长距离的层级关系在所有的自然语言中都存在,这使得短语结构语法成为必不可少(p. 1577)。他们还从人和其他动物比较进化(comparative evolution)的角度推测,人类的递归运算能力不一定是为了语言而发展出来的,而极有可能是为了数数、航行和社会关系等非交际原因。可能源于特定的自然选择的压力、人类所独具的进化历史、或者是其他种类的神经重新组织的后果(副产品)(p. 1578)。

不管怎么说,自然语言的递归性是基于统计的语言处理模型的头号敌人。

6.2 语言学知识的结构依赖性

归根结底,这都是由语言学知识的结构依赖性(structure-dependent)特点造成的。语言学知识的结构依赖性特点,是它不同于人类其他知识的地方。正是根据这一点,乔姆斯基强调语言能力不同于人类的其他认知能力。Chomsky (1980)非常直观地用英语是非问句(yes-or-no question)的形成过程,来说明语言学规则具有结构依赖的特性。例如:

(1) The man is here. —Is the man here?

The man will leave. —Will the man leave?

当考察了上述范围极其有限的陈述句和疑问句配对(declarative-question pair)后,我们或许可以提出下列两种假设(hypotheses)来解释怎样从陈述句上推导出疑问句:

H_1:在陈述句中自左向右逐词搜索,直到发现首先出现的is, will一类词;然后把它放到句首,就形成了相应的疑问句。

H_2:在陈述句中自左向右逐词搜索,直到发现首先出现在第一个名词短语之后的is, will一类词;然后把它放到句首,就形成了相应的疑问句。

像 H_1 这种假设，可以称为“跟结构无关的规则”（structure-independent rule）；像 H_2 这种假设，可以称为“依赖于结构的规则”（structure-dependent rule）。因为，H_1 只需把陈述句分析为一个词的序列（即一连串词，a sequence of words）；而 H_2 除了需要把陈述句分析为一个词的序列外，还需要把陈述句分析为名词短语之类的抽象的短语。之所以说短语是抽象的，是因为它们只是一种心理结构（mental constructions），它们的边界（boundaries）和类别标定（labeling）并不通过某种方式用一般的物理形式标记出来。尽管如此，人们还是愿意选择像 H_2 这种假定了抽象的心智加工过程（abstract mental processing）的假设；因为，它比 H_1 更接近事实。比如，假如有下列陈述句：

(2) The man who is here is tall.
　　The man who is tall will leave.

根据假设 H_2，可以得出下列正确的疑问句：

(3) Is the man who is here tall?
　　Will the man who is tall leave?

但是，根据假设 H_1，却得出了下列不合格的疑问句：

(4) * Is the man who here is tall?
　　* Is the man who tall will leave?

换句话说，假设 H_2 正确地预测了(2)和(3)之间的语法联系，而假设 H_1 根本做不到。通过这些例子，乔姆斯基企图证明：儿童是怎样知道 H_2 是接近正确的呢？显然没有人告诉过他 H_1 不对、H_2 正确之类的相关证据。其实，儿童不需要考虑 H_1 之类的假设，他的大脑的初始状态（initial mental state）的特性一开始就排除了 H_1 这种跟结构无关的规则。①

6.3 花园幽径句挑战概率语法

我们对乔姆斯基这种语言天赋的结论不感兴趣，本文关心的是

① 详见 Chomsky (1980)，p. 39—40.

基于统计的语言模型能否获得在根本上是结构依赖的结构化的语言知识。情况看上去并不乐观。比如，自然语言中有一种在句子的局部有歧义、但整个句子没有歧义的句子。例如：

(1) a. The horse raced past the barn fell.（跑过饲料房的马倒下了。）

b. The student forgot the solution was in the back of the book.

（学生忘记了答案在这本书的背面。）

c. The complex houses married and single students and their families.

（综合建筑物中住着结婚的和独身的大学生以及他们的家庭。）

当人们读到(1a)的前一段 The horse raced past the barn（马跑过了饲料房）时，一般会以为这已经是一个完整的句子了。再往下读到另外一个动词 fell 时，才发觉 raced 原来并不是句子中的主要动词，它是修饰名词 horse 的定语从句中的主要动词，最后读到的 fell 才是这个句子的主要动词。人们对这种句子的理解方式，犹如漫步在花园中曲折的幽径上，出口在哪儿并不是一目了然的。据此，Bever (1970)形象地称这种句子为花园幽径句(garden path sentence)。后来，Trueswell (et al.) (1993)讨论了(1b)这样的花园幽径句。显然，(1a)跟(1b)相对称，前者是主语中包含不带 that、which 类关系代词的定语从句，使人误以为主语中的核心名词跟从句中的动词性成分是一个完整的主谓结构；后者是宾语中包含不带 that、which 类关系代词的定语从句，使人误以为宾语中的核心名词跟主句中的动词性成分是一个完整的述宾结构。而(1c)在理解上的回溯(back-tracking)，则是由 complex 和 houses 的词性歧义造成的。在读到 The complex houses 时，人们一般以为这是一个名词词组，complex 是形容词，它修饰名词 house。但是，当继续向前读到 married and single 的时候，会感到非常迷惘，不明白究竟是什么意思；最后读到句子末尾的时候，才恍然大悟：complex 不应该理解为形容词，而应

该理解为名词;house也不应该理解为名词,而应该理解为动词。这时,整个句子的意思才真正明白。[①]

冯志伟、许福吉(2002)指出,汉语中也有类似的花园幽径句;特别是当潜在的歧义结构(即歧义格式)在实例化(instanciation,即填入具体的词汇)过程中变成了现实的歧义结构后;如果把它们嵌入更大的结构中,那么往往会造成花园幽径句。例如:

(2) a. 小王研究鲁迅的文章 → 小王研究鲁迅的文章发表了。

b. 咬死了猎人的狗 → 咬死了猎人的狗逃跑了。

有的句法结构本来没有歧义,但当后面出现新的成分后,原有的结构要发现变化,并导致语义解释的改变。例如:

(3) 老张讨厌小王 → 老张讨厌小王(的)不老实。

甚至后面出现新的成分后,原有的结构格局可以保持,但语义解释改变并导致语义怪异。例如:

(4) a. 王冕死了 → 王冕死了父亲。

b. 中国队打败了→ 中国队打败了科威特队。

他们认为,汉语的(2)和(3)分别跟英语的(1a)和(1b)相似;但是,像(4)这样的语义花园幽径句是英语所没有的。

对于花园幽径句的歧义段的分析,由于人们一般总是先选择优先的结构(如把"小王研究鲁迅的文章"理解为主谓结构);直到句子快结束时(如读完"小王研究鲁迅的文章发表了"),才发觉非优先的结构(作为偏正结构的"小王研究鲁迅的文章")才是正确的结构。同样,在计算机自动分析这类句子的过程中,往往会出现大量的回溯,从而影响了自动分析的效率。可喜的是,冯志伟、许福吉(2002)报导:他们基于上下文无关语法规则,采用欧雷算法(Earley Algo-

① 本段内容,详见冯志伟、许福吉(2002),第1—3页。

rithm)，[①]成功地处理了英语和汉语中的花园幽径句。由于欧雷算法使用点规则(dotted rule)，把自顶向下(top-down)的"预示"(predictor)和自底向上(bottom-up)的"扫描"(scanner)很巧妙地结合起来；因而用以处理花园幽径句时，完全避免了回溯，提高了分析的效率。在这种实践的基础上，他们对概率语法有如下中肯的评论：

> 近年来，在大规模真实文本的自动处理中，有的学者提出了概率上下文无关文法，他们主张从经过加工的语料库(树库)中统计出上下文无关文法规则的出现概率，然后在语言自动分析中，根据规则概率的大小来优选概率大的规则，从而减少分析过程中对于小概率规则的搜索操作，提高文法的分析效率。
>
> 在上下文无关文法中引入概率的因素，在绝大多数情况下，无疑是有积极作用的。这是自然语言研究中的一个很大的进展。但是，当我们分析花园幽径句的时候，概率语法却会遇到困难。这时，由于概率大的规则反而是分析中不正确的选择，而那些概率小的规则恰恰是分析中的正确选择，使用概率语法会感到束手无策。因此，在花园幽径句的自动分析中，"概率"起的作用是非常特殊的。这说明，概率语法并不是所向无敌的，也不是万能的。花园幽径句的存在，是对概率语法的一个挑战。在自然语言的自动处理中，我们应该看到概率语法的这种局限性，把基于统计的方法和基于规则的方法紧密地结合起来，才能克服概率语法的这种局限性。(第21页)

还是应了一句老话：语言现象比语言理论丰富。同样，语言现象比任何一种语言处理方法或模型都要来得复杂。因此，想单凭一种方法或模型来独打天下的想法是不切实际的。

---

① 关于欧雷算法，详见 Earley (1970)；翁富良、王野翊(1998) § 5.2："上下文无关语法的识别和分析算法"之三"欧雷算法"中有简要的介绍，第69—70页；冯志伟、许福吉(2002)一文中有具体的解释和运用。

7　基于统计的语言处理模型的局限性

7.1　独立性假设：统计语言模型的双刃剑

上面从语言学理论上说明，基于统计的语言处理模型在根本上会碰到不可克服的困难。下面，我们从统计模型内部说明这种方法的局限性。根据§1的讨论，自然语言不是有限状态语言，自然语言的语句中的符号串不是一种马尔科夫链。这样，符号串中某个当前符号的出现概率并不是单纯地由前一个符号决定的，甚至在理论上无法统一地知道到底是由其前的多少个符号决定的。但是，统计模型必须假定当前符号的出现概率是由其前的多少个符号决定的，这就是§5.2中的N元语法模型。这里，引入了概率论上的独立性假设：假定N+1个符号出现这个事件的机会只跟其前的N个符号的出现相关，但是跟语句中的其他符号的出现与否都无关。这已经跟语言的实际情况相对立了。并且，在实际构造和实现统计模型的时候，为了避免自由参数太多而造成的计算上的指数爆炸，同时为了克服训练例中数据稀疏的困难；这个N的数目不能太大，通常要减少到3以下才能实施。这样，势必使得这种基于统计的语言处理模型离语言事实越来越远了。

简单地说，独立性假设是一把双刃剑：基于统计的语言处理模型借助于独立性假设，使得统计模型得以实施；但是，独立性假设过度地简化了语言模型，使得统计模型只能处理对结构关系依赖性不强的对象，而像代词的先行词求解、长距离依存关系等依赖结构关系的结构化对象，则较难用统计模型来处理。不幸的是，绝大部分语言学知识和语法规则都具有结构依赖的性质，它们使得独立性假设失效，从而使得统计模型难以施展神威。

7.2　一个实验：介词结构消歧的条件

据黄昌宁（2002：83）介绍，Collins and Brooks（1995）用§5.8中的概率统计方法来进行介词结构消歧实验。他们采用宾州大学提

供的带有句法标注的华尔街日报(WSJ)树库,从中抽出 20,801 个四元组作为训练集,其余的 3,097 个四元组作为测试集。并把机器自动判定的结果跟人在不同条件下的判定结果进行比较:

实验条件	精确率
一律视为名词附加(即 $A \equiv NP$)	59.0%
只考虑句中介词 p 的最常见附加	72.2%
机器根据四个中心词判断	84.1%
三位专家只根据四个中心词判断	88.2%
三位专家根据全句判断	93.2%

显然,自动判断的精确率的下限是 72.2%,因为机器根据四个中心词来判断,不会比只考虑句中介词 p 的最常见附加做得更差;上限是 88.2%,因为同样根据四个中心词来判断,机器不会比专家更高明。虽然自动判断的精确率 84.1%离上限 88.2%不远,但是离实际应用的需要还有距离。究其原因,语法规则的结构依赖性未必是 N 元语法模型所能逼近的,更何况除了语言知识之外,语篇上下文知识和世界知识对语言理解也产生影响;因此,即使是人类专家、即使阅读了全句,也未必能正确地判断出介词结构的正确附加。语言自动处理的困难性,由此可见一斑。

7.3 一点疑问:语义等价类能管多大用处?

从上面介词结构消歧的例子来看,不管是人还是机器,为了达到 80%以上的正确率,就得考虑四个中心词,即使用四元语法模型。但是,这就马上带来计算量大和数据稀疏的问题。为了解决这个问题,可以采用划分等价类的办法,把名词和动词按其语义划分成若干等价类,从而把相关于一个个具体的名词和动词的介词结构附加模型转变成相关于这些语义等价类的模型。比如,把 Monday, today, March 划归等价类 TIME,把 John, baby, boy, girl, artist 划归等价类 HUMAN。这样,介词结构附加模型不是相关于具体的名词和动词,而是相关于这些词所属的类。即

$$P(A \mid prep, verb, noun_1, noun_2) =$$

$$P(A \mid prep, c(verb), c(noun_1), c(noun_2)) \quad (7.3.1)$$

其中,$c(X)$是词 X 所属的语义等价类。由于等价类的数目远远小于具体的名词和动词的数目,因而公式(7.3.1)大大地减少了自由参数的数量。如果进一步假设:

$$P(NP \mid prep, c(verb), c(noun_1), c(noun_2)) > P(VB \mid prep, c(verb), c(noun_1), c(noun_2))$$
$$\Updownarrow$$
$$P(c(noun_1), c(noun_2) \mid prep) > P(c(verb), c(noun_2) \mid prep) \quad (7.3.2)$$

也就是说,简单地比较在特定介词的情况下,介词后的名词所属的语义等价类,到底跟中心语名词所属的语义等价类的共现概率高,还是跟动词所属的语义等价类的共现概率高,来估计哪一种附加的可能性更大。这里的 $P(c(noun_1), c(noun_2) \mid prep)$ 和 $P(c(verb), c(noun_2) \mid prep)$,可以根据最大似然估计原理来求得。同样可以采用§5.8 所述的办法,从非歧义的数据中得到最大似然估计所需要的有关事件的计数。[①]

这里事实上引进了一个假设:相同语义类的词具有相同的语法表现(grammatical behavior)。这是一个有待验证的假设。根据我们初步的经验,相同语义类的词不一定具有相同的语法表现;因此,它们未必可以划入相同的词类。汉语中有名的例子是:

打仗(动词)~战争(名词)　金、银(区别词)~铜、铁、锡(名词)

绿(形容词)~碧绿(区别词)　红(形容词)~通红(状态词)

突然(形容词)~忽然(副词)　刚才(时间词)~刚刚(副词)

属于相同意义范畴的动词,不一定能构成相同的句式。例如:[②]

(1) a. Joe gave $5 to the earthquake relief fund.

→ b. Joe gave the earthquake relief fund $5.

---

① 详见翁富良、王野翊(1998),第 177—180 页。

② 例子和说明,根据 Goldberg (1995),p. 121、130—131 改编。

(2) a. Joe donated $5 to the earthquake relief fund.
→ b. * Joe donated the earthquake relief fund $5.
(3) a. Joe told the news to Mary.
→ b. Joe told Mary the news.
(4) a. Joe whispered the news to Mary.
→ b. * Joe whispered Mary the news.
(5) a. Joe baked a cake for Mary.
→ b. Joe baked Mary a cake.
(6) a. Joe iced a cake for Mary.
→ b. * Joe iced Mary a cake.
(7) a. She threw a cannonball to him.
→ b. She threw him a cannonball.
(8) a. She blasted a cannonball to him.
→ b. * She blasted him a cannonball.
(9) Sally permitted/allowed/ * let/ * enabled Bob a kiss.
(10) Sally refused/denied/ * prevented/ * disallowed/ * forbade Bob a kiss.

从例(1)—(8)可以看出，同样是给予义动词，give可以进入双宾句式、而donate不行；同样是言说义动词，tell可以、而whisper不行；同样是制作(creation)义动词，bake可以、而ice不行；同样是弹道运动(ballistic motion)义动词，threw可以、而blast不行。从例(9)和(10)可以看出，同样是许可(permission)义动词，permit，allow可以、而let，enable不行；同样是拒绝(refusal)义动词，refused，deny可以、而prevented，disallow，forbid不行。汉语的情况也一样，例如：

(11) a. 我吃了弟弟一个苹果
b. * 我啃了弟弟一个猪手
c. * 我嚼了弟弟一根香蕉
d. * 我尝了弟弟一口蛋汤
(12) a. 我穿过舅舅一件毛衣

b. 我戴过舅舅一顶帽子

c. ＊我披过舅舅一件斗篷

d. ＊我围过舅舅一条纱巾

(13) a. 动物园飞了一只鹦鹉

b. ＊动物园蹿了一只豹子

c. ＊动物园蹦了一只袋鼠

d. ＊动物园跳了一只猴子

e. ＊动物园溜了一只狐狸

f. ＊动物园走了一只孔雀

g. ＊动物园滚了一只猪獾

h. ＊动物园爬了一只乌龟

i. ＊动物园游了一只白鹅

同样是二价的摄食动词，“吃”可以进入三价句式，但“啃、嚼、尝”不能；同样是二价的服饰动词，“穿、戴”可以进入三价句式，但“披、围”不能；同样是一价的移动动词，“飞”可以进入二价句式，但“蹿、蹦、跳、溜、走、滚、爬、游”不能。可见，语义上的等价类，不一定是句法上的同分布类。

8　结语：走向统计方法和规则方法的结合

根据上文的讨论，线性的语法模式难以处理语言中的嵌套结构。在目前的技术条件下，基于统计的语言处理模型无法通过对线性的语言符号序列上有限的N个符号之间共现概率的统计，来发现真正的语法结构、从而达到真正的语义理解。对此，黄昌宁、李涓子(2002)有过极为精到的表述：

> ……自然语言最重要的特征是其结构性，而N元语法模型是一种基于线性的符号同现关系的语言模型，只能观察到表示语言最表层信息的符号（一般为字、词或词性标记）之间相邻出现的现象，并不能观察语言的结构，因此用线性的N元语法模型来表示结构化的自然语言具有局限性。

……HMM 等价于概率型正规文法，是一种有限状态模型，而有限状态不能描述自然语言的层级结构。（第 118—119 页）

这倒使人想起将近半个世纪前，Chomsky（1957：16—17）的断言：

一个人说出和理解符合语法的话语的能力，并不是建立在统计逼近（statistical approximation）之类的概念基础上的。……语法学是自成一系的（autonomous），是离开语义而独立的；概率论模式（probabilistic models）无助于人们彻底理解句法结构上的一些问题。（中译本，第 10—11 页）

随着语言分析技术和计算技术的进步，乔氏的断言不但没有成为过时的教条，反而是不幸而言中。这使得我们思考：语言信息处理面临的对象既然有如此顽劣的既抗拒规则模型、又抗拒统计模型的属性，那么一种可能的技术途径只能是把规则的方法和统计的方法结合起来，采用多元化的方法来建立处理自然语言这种混杂（miscellaneous）系统的综合性模型。因此，不管是追求规则挖掘的语言学家、还是沉迷概率统计的计算语言学家，对于语言信息处理，大家面前都有许多紧迫的工作值得去做。

鸣谢：本文承詹卫东先生指正并提供有关资料和技术支持，谨此致以诚挚的谢意。

参考文献

白栓虎（1992）《基于统计的汉语语料库词性自动标注的研究与实现》，清华大学计算机系硕士论文。收入黄昌宁、夏莹主编（1996）《语言信息处理专论》，清华大学出版社，第 37—77 页。

范继淹（1964/1983）《汉语语法结构的层次分析问题》，《语法研究和探索》第 1 辑，第 57—84 页，北京大学出版社。收入《范继淹语言学论文集》，第 211—238 页，语文出版社，1986 年。

冯志伟、许福吉（2002）《花园幽径句初探》，The 2nd Kent Ridge International Roundtable Conference on Chinese Linguistics（theme：Syntax-Morphology-Phonology Interface），National University of Singapore，November 27—29，2002.

石纯一、黄昌宁、王家廞 (1993)《人工智能原理》,清华大学出版社。

黄昌宁 (2002)《统计语言模型能做什么?》,《语言文字应用》第1期,第77—84页。

黄昌宁、李涓子 (2002)《语料库语言学》,商务印书馆。

翁富良、王野翊 (1998)《计算语言学导论》,中国社会科学出版社。

张立昂 (1996)《可计算性与计算复杂性导引》,北京大学出版社。

朱德熙 (1980)《汉语句法中的歧义现象》,《中国语文》第2期。收入朱德熙(1980)《现代汉语语法研究》,商务印书馆。

Bever, T. G. (1970) The Cognitive Basis for Linguistic Structures. In J. R. Hayes (1970) (ed.) *Cognition and Development of Language*, p. 279—352, New York: Wiley.

Bloomfield, L. (1935) *Language.* Allen & Unwin.

Chatman, Seymour (1955) Immediate Constructions and Expansion Analysis. *Word*, Vol. 11, p. 377—85.

Chomsky, Noam (1956) "Three Models for the Description of Language." *IRE Transactions on Information Theory*, IT—2, p. 113—124, Proceedings of the symposium on information theory, Sept., 1956.

Chomsky, Noam (1957) *Syntactic Structure*. The Hague: Mouton.《句法结构》,邢公畹、庞秉均、黄长著、林书武译,中国社会科学出版社,1979年。

Chomsky, Noam (1964a) *Current Issues in Linguistic Theory*. The Hague: Mouton.

Chomsky, Noam (1964b) The Logic Basis of Linguistic Theory. In *Proceedings of the Ninth International Congress of Linguists*, edited by H. Lunt, pp. 914—978. The Hague: Mouton. (A revised version was published as Chomsky 1964a).

Chomsky, Noam (1965) *Aspects of Syntactic Theory*. The Hague: Mouton.

Chomsky, Noam (1980) On Cognitive Structures and Their Development: A Reply to Piaget. In Piattelli-Palmarini, Massimo (1980) (ed.) *Language and Learning: The Debate between Jean Piaget and Noam Chomsky*, p. 35—54. Cambridge: Harvard University Press.

Chomsky, Noam & Miller, George (1963) Introduction to the Formal Analysis of Natural Languages. In *Handbook of Mathematical Psychology*, edited by p. Luce, R. Bush, and E. Galanter, Vol. II, pp. 269—322. New York: Wiley.

Collins, M & J. Brooks (1995) Preposition Phrase Attachment through a Backed-off Model. in *Proceedings of the 3rd WVLC*, Cambridge, MA.

Corder, S. Pit (1979) *Introducing Applied Linguistics*. Penguin.《应用语言学导论》,上海外国语学院外国语言文学研究所译,上海外语教育出版社,1983 年。

Earley, J. (1970) An Efficient Context-free Parsing Algorithm. CACM 13 (2), p. 94—102.

Freedman, David, Robert Pisani, Roger Purves & Ani Adhikari (1991) *Statistics*, second edition. W. W. Norton & Company, Inc.《统计学》,魏宗舒、施锡铨、林举干、李毅译,吕乃刚、范正绮、吴喜之校,中国统计出版社,1997 年。

Fries, C. C. (1957) *The Structure of English: An Introduction to the Construction of English Sentences*. London: Longman.《英语结构——英语句子构造导论》,何乐士、金有景、邵荣芬、刘坚、范继淹译,范继淹、金有景校订,商务印书馆,1964 年。

Garside, R., G. Leech, & G. Sampson (1989) (ed) *The Computational Analysis of English: A Corpus-Based Approach*. London: Longman.

Goldberg, E. Adele (1995) *Constructions: A Construction Grammar Approach to Argument Structure*. Chicago and London: The University of Chicago Press.

Hauser, Marc D., Norm Chomsky & W. Tecumseh Fitch (2002) The Faculty of Language: what is it, and how did it evolve? *Science* 298, p. 1569—1579.

Hockett, F. Charles (1953) Reviews of C. Shannon & W. Weaver, *The Mathematical Theory of Communication*. *Language*, 29, p. 69—93.

Hockett, F. Charles (1955) *A Manual of Phonology*. Baltimore: Waverly Press.

Hockett, F. Charles (1958) *A Course in Modern Linguistics*. Macmillan Publishing Co., INC.《现代语言学教程》,索振羽、叶蜚声译,北京大学出版社,2002 年。

Newmeyer, J. Frederick (1986) *Linguistic Theory in America*, Second Edition. Orlando: Academic Press, INC.《当代美国语言学史》,吴黉铭译,台北:文鹤出版有限公司,1998 年。

Pittman, S. Richard (1948) Nuclear Structures in Linguistics. *Language*, 24,

p. 287—292; in Joos (1958) (ed), p. 275—78.《语言学中的核心结构》,劳宁译,收入《中国语文》编辑部(1963)《语言学资料》6:描写语言学(语法部分)专号,第68—71页。

Shannon, E., Claude (1948) The Mathematical Theory of Communication, *Bell System Technical Journal*, July and October, 1948. Reprinted in Shannon and Weaver (1949) p. 3—91.

Shannon, E., Claude & Warren Weaver (1949) *The Mathematical Theory of Communication*. University of Illinois Press.

Trueswell, J. C. (et al.) (1993) Verb-specific Constraints in Sentence Processing: Separating Effects of Lexical Preference from Garden-paths. *Journal of Experimental Psychology: Learning, Memory and Cognition*, 19 (3), p. 528—553.

Weaver, Warren (1949) Recent Contributions to the Mathematical Theory of Communication, in Shannon and Weaver(1949)p. 93—117. Its condensation previously appeared in *Scientific American*, July, 1949.

2002年12月初稿,2003年3月改定

(删节发表于《语言文字应用》2004年第2期)

认知科学和汉语计算语言学

本文讨论认知科学跟汉语计算语言学的关系。第一部分讨论语言研究怎样以当代科技,特别是以计算机科学和技术为参照,揭示语言结构和意义的有关规律;第二部分讨论智能系统和认知科学的关系,特别是认知科学的研究内容、基本假设和方法论特点;第三部分讨论认知科学跟语言学的相互影响,特别是语言学研究对认知科学的贡献;第四部分讨论认知科学和计算机理解自然语言研究的关系,包括怎样从对语言的认知研究走向对语言的计算分析,特别讨论了基于认知并面向计算的汉语语法研究的路线;第五部分简单介绍从事语言学学习和研究的人怎样逐步进入计算语言学研究领域。

1 语言研究的价值取向和评价参照

为什么要研究语言?语言学有什么用?怎样来评价语言学的研究成果及其所达到的水平?这是许多语言学家,特别是初涉语言学的学生爱问的问题。下面就这三个问题略作讨论,作为本文的一个引子。

1.1 语言学的研究空间

语言学有两种定义方式,一种是传统的,即语言学是研究语言的科学;一种是现代的,即语言学是对语言的科学研究。① 这两个定义都涉及到"语言"和"科学"这两个关键词。但是,语言的范围很广泛,从语音、词汇、句法、语义一直到语用;科学研究的范式(paradigm)很多,从结构主义到后结构主义、从功能主义到形式主义。可见,语言学的研究空间太大。对此,每个研究者都必须作出选择。而选择的

① Lyons (1968: 1)说: Linguistics may be defined as the scientific study of language.

依据在很大程度上取决于研究者的价值观念,即为了什么而研究语言。

1.2 面向当代科技的语言研究

在众多的语言学追求中,我们倡导一种面向当代科技的语言研究,强调语言研究的当代性和应用性。具体地说,包括下面两层意思:(1) 语言研究为当代科学技术服务,使语言学的研究成果更具有科学的认识价值和实际的应用价值。比如,对于人类自身智能的认识,可以从语言角度切入,从而开拓新的思路:模拟大脑的活动而发展新的计算机原理、新的计算方法和软件技术。对语言结构的精细的描写和形式化处理,可以为计算机处理自然语言提供可靠的基础,从而开辟语言信息处理产业这一新的市场。

(2) 用当代科学技术的新观念来冲击、刷新语言学的理论和方法。比如,Chomsky 的生成语法就是在上个世纪 50 年代的计算机科学技术、认知科学、数理逻辑、信息论等当时最新的科学技术的背景上产生的。同时,反过来又对当时的科学技术,比如理论计算机科学(特别是形式语言的层级体系,即 Chomsky Hierarchy),起了促进作用。

1.3 语言研究的计算机参照

怎样来评价一个时代的语言研究的成就和所达到的水平呢?白硕(1996)认为:评价语言学知识需要参照物作为"硬"的检验。比如,传统语言学以本族说话人为参照物,以满足本族语言教学的需要为目的。因此,所获得的语言知识在今天看来不完善和过于简化。描写语言学以非本族说话人为参照物,以满足外语教学和对异文化的了解为目的(比如,上世纪初人类学家对各种印第安语言的记录和描写)。许多本族人习以为常的现象被挖掘出来了,语言学知识从量到质都有了明显的提高。后来,出现了计算机和计算机理解自然语言,于是,计算机成了语言学知识的一个新的参照物。因为计算机只能处理形式化的知识,所以要想让计算机处理自然语言,就必须把语言学知识形式化。正是在把语言学知识形式化的过程中,人们认识

到了一些没有计算机作参照就很难揭示出来的现象和规律。现在，网络将成为语言学知识的一个新的参照物。因为在网上传输的信息很大一部分是自然语言，所以语言学必然要在网络信息处理中扮演重要的角色。比如，网络信息的文本分类、快速检索、信息抽取、信息过滤等，都需要语言学知识作支持。①

这样，语言研究就不仅具备自然科学的探索、认识功能，而且还具备技术科学的社会功能——利用对语言的科学认识来造福于人类。

2 智能系统和认知科学

语言是人类智能的重要的组成部分，而新兴的认知科学以研究智能系统为己任。因此，认知科学势必会对语言学产生积极的影响，并为语言学的科学化和现代化提供机会。为此，下面对认知科学中若干重要的方面略作介绍。

2.1 心脑的二元对立和认知中介理论

众所周知，人是一种有心智的动物。所谓心智(mind)泛指人的知觉、注意、记忆、学习、思维、理解、创新等各种心理活动，它跟大脑(brain)相对。在人类的心智中，像判断、推理和想象等利用知识去解决问题的心理能力被称为智能(intelligence)。智能也可以定义为在新情况下作出恰当的反应的能力，因为要在新的情况下作出恰当的反应，必然要利用知识来进行判断、推理和想象。至于大脑则是心智的器官，大脑的活动(即脑过程)其结果产生了心智。脑过程表现为大脑中的神经元(neuron)之间传递信息的生物电学和化学过程。

问题是：如何用大脑中的神经元的活动这种低层次的生理现象去说明、解释心智这种高层次的心理现象。为了填补这种心脑二元之间的鸿沟，功能主义者假设在人的大脑和心智之间存在着一个认

① 笔者在引述时作了补充和发挥，如有差错，责任在我。

知平面，在这个抽象的理论平面上，我们可以撇开脑过程这种具体的生化现象来谈论大脑是怎样工作的。可见，认知(cognition)是功能主义者对人类智能在大脑中的组织方式和工作原理的一种理论概括，它包括认知结构(意象〔imagery〕、图式〔scheme〕、范畴、原型、命题、脚本、网络等)和认知过程(如记忆、编码、搜索、思维、概念形成、扩散性激活、缺省推理、隐喻投射、语言理解等)两个方面。

2.2 什么是认知(活动)?

认知有时指认知活动，在这一意义上，认知指人用知识去解决复杂问题的心理过程。[1] 认知活动一般不包括像感觉、知觉等低层次的心理活动。比如，对光点的感觉、图形知觉的形成，一般来说不属于认知活动，因为它并不利用知识。但是，当人们把北极星周围的一群星看作一只小熊(命名为小熊星座)，把其附近的另一群星看作一只大熊(命名为大熊星座)，并把大熊星座的七颗明亮的分布成勺形的星看作盛酒的斗(命名为北斗星)；那就属于认知活动，因为这是一种基于知识的隐喻投射——把人们生活中熟悉的概念投射到陌生的事物上。认知活动通常指高层次的心理活动，如问题求解(problem resolution)，像求解代数方程式等活动。比如，已知方程式：$8x+5=4x+17$，求解 $x=$?

要求解出 x 的值，必须对给定的方程式进行变换，最后得到 $x=\cdots$这样的形式。在这过程中，必须遵循这样的等价变换规则(rule)：在方程式的等号两边同时加、减、乘、除相同的数，等式不变。其实，规则只是一种约束条件，人们还必须使用策略(strategy)来作宏观的指导，以明确什么时候、什么情况下使用什么规则、进行什么操作。在问题求解过程中，最常用和有效的策略是“手段—目的”分析法(mean-end analysis)。比如，在解上列方程时，为了达到求得 $x=\cdots$这样的目的，得设法消去等式右侧的4x和等式左侧的常数5和系数8。于是，在等式两边同时减去4x，减去5，再除以8，就得到了 $x=\frac{3}{2}$。事实上，人们已经把这种策略和规则结合在一起，总结成程式化的口

---

① §2.2—2.6参考李家治(1985)等文献，不一一具指。

诀：移项合并同类项。

虽然，这种问题求解是一种非常复杂的认知活动；但是，通过分析其过程及其所使用的策略和规则，可以写出极其机械的形式化的解题方法，即算法（algorithm）。比如：

if "X＝N"→ Hold & check； N：number（数）
if N on left → S(N)； S：subtract（减）
if Nx on right → S(Nx)；
if Nx on left，N≠1 → D(N). D：divide（除）

如果用某种程序语言来把上述算法编成程序，那么就可以在计算机上运转，即进行自动解方程。这个例子说明，对人类的心理活动的认知研究，最终可以导向一种非常严格的计算分析。或者说，认知的本质是计算，表现为一系列受约束的变换操作；其中的每一步都是由目标制导的（goal-directed），并且是受规则约束的（rule-constrained）。这一点，下一节还要讨论到。

2.3 什么是认知科学？

简单地说，认知科学（cognitive science）是研究心智的科学；具体地说，认知科学是一门研究智能系统（包括天然的和人工的）的内部结构、功能和工作原理的科学。这里，天然的智能系统指人的大脑，人工的智能系统指计算机。认知科学是一门新兴的前沿性学科，它是在哲学、心理学、语言学、计算机科学和神经生理学等多个学科的交叉领域中发展起来的。

认知科学用信息加工（information processing）的观点来研究认知结构和认知过程，比如，把记忆比作计算机的存储器、把思维比作信息加工（即对符号串进行受约束的变换）等。像 H. Simon 和 A. Newell 还提出了著名的物理符号系统假设（hypothesis of physical symbolic system）：智能的基础是符号操作，通过符号的产生、排列和组合，智能系统就能将外部的事件内化为内部的符号事件并加以控制，从而表现出智能来。因此，一切认知系统（不管是天然的人脑还是人工的电脑）的本质都是符号加工系统。而符号操作的实质就

是计算(computation),表现为具有特定语义解释的符号表达式的各种受规则约束的变换。比如,人的心智表达就是一种形式化的符号表达式,是跟系统的物理状态(即神经元的某种运动方式)相对应的某些基本要素的离散的排列。所有跟系统有关的语义内容都依靠深层的符号表达式及其变换形式和符号关系结构来规定。显然,这是一种语义上中断的物理符号操作,因而是一种计算。因此,"认知就是计算"是经典的认知科学的一个信条。

2.4 认知科学的历史背景

人类的智能问题一直是哲学家关心的话题,从柏拉图到笛卡儿等伟大的哲学家对此都有过精辟的论述。但是,直到计算机出现,并且涉及到计算机模拟人类智能问题时,认知科学这个学科及其特有的性质才得以确立。

1956 年在 MIT 召开了关于通讯和信息论的学术会议,心理学家 Miller 提交了关于短时记忆的容量的论文,心理学家 Bruner 提交了关于思维研究的论文,语言学家 Chomsky 提交了关于语法的形式特性的论文,计算机科学家 A. Newell、心理学家 H. Simon 提交了关于"逻辑理论家"的论文(旨在使计算机可以使用启发式(heuristic)程序像人一样解决问题)。这种对智能系统的多学科的合作和交流,使得认知科学初具雏形。同年,计算机科学家 M. Minsky、J. McCarthy、A. Newell 和心理学家 H. Simon 等聚首普利茅茨学院,探讨一些计算机科学技术方面的问题。他们特别讨论到了用计算机来模拟人类智能的问题,McCarthy 还专门造了 artificial intelligence (人工智能)这一名词。用计算机模拟人类智能的思想又推动了认知科学的产生。

1975 年 Chomsky 和心理学家 J. Piaget 关于人类智能的来源当面进行辩论。[①] 心理学家 Gardner 对此事评论时宣称 Cognition comes of age(认知的时代到来了)。1977 年 *Cognitive Science*(认知科学)杂志创刊,成立认知科学学会,并以该杂志为会刊;1979 年召

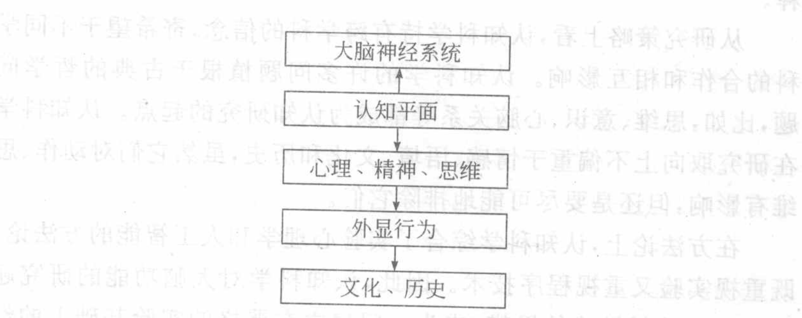
① 详细的情况,请看 Piattelli-Palmarini (1980) (ed.)。

开认知科学学会的第一次正式的年会，这都标志着认知科学的诞生。

2.5 认知科学的研究内容和核心假设

认知科学的特点是：范围广泛、核心明确、层级清晰。凡是跟心智有关的问题，从神经基础到社会文化因素、从哲学思辨到计算机程序实现，都是认知科学家所津津乐道的。他们经常讨论的问题有：(1) 复杂行为的神经生理基础、遗传因素；(2) 问题求解和推理过程；(3) 符号系统，包括自然语言、语音、图像、数字、视觉映象等；(4) 知觉的呈现(presentation)和符号表征(representation)问题；(5) 记忆模型，如工作记忆、短时记忆、中期记忆、长时记忆等；(6) 知识表示理论，如心理表象(即意象)、图式、范畴、原型、命题、框架、脚本、网络等；(7) 自然语言的理解和生成；(8) 学习的模型，涉及问题的表示、解题的条件和动作等；(9) 目的、情绪、动机对认知的影响；(10) 社会文化背景对认知的影响。

这样，构成了认知科学的如下这种以认知平面为核心的研究层次：



至于为什么要研究这些内容以及怎样来研究这些纷繁的问题，认知科学基于如下两个重要的核心假设：

(1) 存在着认知这一独立的心理表示平面。认知科学认为人类的认知活动必须用符号、图式(schemes)、表象(imagery)、观念(idea)和其他心理表示形式来加以描述。在这样的表示平面上进行研究时，科学家处理的是像符号、规则、表象之类用以表示事物的实体，这种实体是处于输入和输出之间的表示材料；据此，可以探索连

接、转换或比较这些表示实体的方式。为了解释各种各样的人类行为、动作和思维,这种水平是十分必要的。

(2) 可以把计算机作为人类思维的模型。如果说计算机有转换、处理信息、进行推理、改变行为的能力,那么可以用同样的方式来刻画人类的思维特征,也完全可以用计算机来模拟人类的认知过程。这就是2.3中提到的物理符号系统假设。

2.6 认知科学的学科性质、研究策略、方法论特点

从学科性质上看,认知科学是一门新兴的交叉科学,它跨接心理学、脑神经科学、计算机科学和语言学等多种学科。虽然认知科学的各种来源学科有共同的目标:发现心智的表示和计算能力及其在人脑结构和功能上的表现,并在计算机上模拟验证。但是,认知科学没有形成公认统一的研究范式,即没有一致的假设和方法。所以,从事不同学科领域的认知科学家倾向于把自己喜爱的范式加之于整个领域,并希望把认知科学理解成符合他们各自对于该领域的尝试性解释。

从研究策略上看,认知科学持有跨学科的信念,寄希望于不同学科的合作和相互影响。认知科学的许多问题植根于古典的哲学问题,比如,思维、意识、心脑关系等都成为认知研究的起点。认知科学在研究取向上不偏重于情感、语境、文化和历史,虽然它们对动作、思维有影响,但还是要尽可能地排除它们。

在方法论上,认知科学综合了实验心理学和人工智能的方法论,既重视实验又重视程序技术。因此,认知科学对大脑功能的研究超越了以往的哲学式的思辨,成为一门建立在严格的实验基础上的经验性的科学。

3 认知科学和语言学的相互影响

认知科学无疑将为语言学提供新的研究范式,同时,语言学也将为认知科学提供广泛而系统的素材和虽然不甚严格但确实是富有洞察力的方法。

3.1 认知科学对语言研究的影响

认知科学可以帮助我们形成新的语言观和方法论。我们应该把语言学置于认知科学的洪流中,使之成为更为广阔的探索人类心智的伟大事业的一部分。这样,可以扩大语言学家的眼界,帮助我们形成新的语言观和方法论。从认知的角度看,语言是人类普遍的认知组织的一个组成部分,它既是认知的工具和手段,又是认知的结果。同时,认知科学的设计实验的方法、建立模型的方法、假设抽象的心理表示平面的方法、用计算机进行模拟和验证的方法,都对语言研究的方法论革新具有特别重要的启迪作用。

认知科学可以帮助我们从人类认知的角度去理解、评价形式语法和功能语法的各种理论模型,特别是其中的语言知识的表示平面和形式化的表示方法,检验这些理论模型中的有关概念、规则、假设等的心理现实性,从而促使我们去建立更加有效的语言学模型。

认知科学还可以推动语言学跻身于当代前沿科学。正如周光召(1995)所言:"人的思维和意识是如何由人脑产生的?能不能用计算机加以模拟?这是最基本的科学问题之一。人的大脑是自然过程中最伟大的杰作,彻底地揭开大脑的奥秘是自然科学面临的最大挑战。……大脑在逻辑运算上虽不如一台高速运行的计算机,但图形识别和直觉判断的能力则远远高于一台超级计算机。这一矛盾暗示我们:人脑的工作原理不同于目前计算机的结构和运算方式,它除了逻辑思维、抽象思维以外,还有形象思维。……因此,探索人脑的认知过程和模式,对创造性地发展计算机科学,是一项必要的选择。认知科学是在神经科学、心理学、科学语言学、计算机科学乃至哲学的交界面上发展起来的,它以人类的智能和认知活动为研究对象。"于是,作为认知科学的一个组成部分的语言学,必将在认知科学的要求和带动下,不断地完善自己;并且,随着认知科学一起跻身于当代前沿科学之林。

3.2 语言学对认知科学的贡献

语言是人类智能的一个重要的组成部分,语言能力是人类最基

本的认知能力。因此,语言是洞察人类心智的一个窗口,研究语言在一定程度上就是在研究心智。比如,认知科学惯于从人类显现的行为来推断有机体的心智能力,再从心智能力来推断有机体的某些性质。于是,通过观察人类的语言行为,便可以推断人类的语言能力;再由此推断人类的某些心智特征(比如递归性能力,等等)。从儿童语言获得(language acquisition)的如下两个事实:速度快、输入极其不完善,可以得出人类有天生的语言能力(linguistic competence)这一假设;再由此可以推出人类大脑的有关机制。加上语言研究的历史长、成果多、结论又相对一致,①这特别有利于语言学成为认知科学的一个核心的组成部分。

另外,语言最有系统性,也最便于观察。于是,我们可以通过分析其输入和输出关系,来假设介于输入和输出两端之间的人脑的工作机制,从而对人脑的语言处理机制作出认知假设。特别是,认知研究在很大程度上是一种理论假设,即所谓的黑箱模型(black box model);这跟神经生理学的解剖实验和实证研究尚有很大的鸿沟。但是,通过对语言研究,特别是对失语症病人的语言缺损情况跟大脑损伤部位的关联性研究,在一定程度上可以把高层次的认知研究跟低层次的神经研究沟通和关联起来;从而,使认知科学的理论假设能够建立在神经科学的实证基础上。为此,下面专门介绍关于语言结构中的空语类(empty category)的心理实验及其跟大脑损伤部位的关系的有关实验。②

3.2.1 语言结构中的空语类及其心理现实性

大家知道,一个句子的意思并不是这个句子中的词语的意思的简单的堆砌;除了词语之外,还有结构要素在其中起作用。看得见的结构要素包括语序(word-order)、形态变化(inflection)和虚词(function word)等,看不见的如结构层次(structural hierarchy)和结构关系(structural relation)等。例如:

(1) a. I know who$_i$ Josephine thinks [e$_i$] is clever.

---

① 参考 Halle (1973)。

② 下面的介绍详见 Fodor (1995)。

b. I know whom$_i$ Josephine ought to consult [e_i].

(2) a. Which books$_i$ did John read [e_i] in the bathtub?

b. Do you recall which books$_i$ John proclaimed [e_i] were unreadable?

从例(1)可以看出,关系代词的形态格(主格还是宾格)的确定取决于其在底层结构(underlying structure)中的位置。从例(2)可以看出,虽然疑问短语(Wh-phrase)都处于从句的句首位置,但是由于其论旨角色(thematic role)不同,因而在句子的语义解释中的地位不同。像这种细微的差别,只能从其潜在的底层位置(underlying position)上去推求和解释。于是,从理论上,或者说从认知上,可以假定其原来的位置上还留下一个语迹(trace),即空语类(可以用 e 来代表)。语迹虽然没有语音形式,但是具有句法作用;它约束前移的疑问短语,并在语义上跟这个前移的成分(即先行成分)同指(co-reference)。作为约定,可以用下标来标注它跟其先行成分的同指关系。

问题是,这种认知假设有没有心理现实性(psychological reality)? 这只有通过心理实验来检验。下面,介绍三个这方面的心理实验。

实验一:自定时间的阅读方式(self-paced reading paradigm)。例如:

(3) a. What$_i$ did the cautious old man whisper [e_i] to his fiancée during the movie last night?

b. What$_i$ did the cautious old man whisper to his fiancée about [e_i] during the movie last night?

被试按一下按钮,屏幕上出现一个词;被试觉得自己理解以后,再按一下按钮,申请下一个词。时间长短不限,但都进入统计。结果,理解(3a)中的 to his fiancée during the movie last night 快于(3b)中的 to his fiancée about [e_i] during the movie last night。因为 whisper 有及物(如 John whisper a message to his friend)和不及物(如 John whisper to his friend,或者 John whisper about the message)两种用

法，所以当被试读到(3b)的 whisper 时，以为那儿跟(3a)一样有一个空语类；后来读到后面的词语才知道空语类原来在 about 之后。正是这种错误的插入空语类以及后来的修正，多花了语义理解的时间。这说明人在处理句子时，不仅能识别空语类的位置，而且有急于找出跟前移成分相关的语迹的心理倾向。

实验二：视觉探测识别(visual probe recognition)。例如：

(4) a. The terrorists wanted to disrupt the ceremonies.

b. [The new mayor at the center podium]$_i$ was shot [e$_i$].

(5) a. The terrorists wanted to disrupt the ceremonies.

b. The new mayor at the center podium was furious.

在屏幕上显示上面的句子 b 及其背景句 a，然后消失；再显示 mayor 等探测词，要求判断它是否在前面的句子中出现。结果，正确判断(4)类句子的时间短于(5)类句子。一种可能的解释是，句尾的空语类跟前面的先行词同指，被激活(activate)的先行词有助于对探测词的判定。

实验三：听觉-视觉交叉模式启动(cross-modal priming)。例如：

(6) The policeman saw the boy$_i$ [that$_i$ the crowd at party accused [e$_i$] of the crime].

让被试听上面的句子，到空语类处[e$_i$]在屏幕上显示 girl 等(跟 boy 相关)词，让被试大声读出。计算从显示到读出所花的时间。也显示 officer(跟 policeman 相关)、people(跟 crowd 相关)，以及其他无关的词作为对照。结果，只有在[e$_i$]处且跟 boy 相关的词反应时间最短。这也说明了空语类的存在，并在句子的语义解释中起作用。

3.2.2 处理空语类的神经基础

上面只是在比较抽象的心理学层面上，证明空语类具有心理现实性。现在的问题是，人类处理空语类的神经基础是什么？由于不能解剖正常人的大脑等伦理限制，只能从失语症患者的语言表现那儿间接地寻找答案。

失语症(aphasia)指大脑一定区域发生器质性病变而造成的言语缺失,即语言表达或理解上的障碍。其中,比较典型的有两种:(1)布洛卡失语症(Broca's aphasia),表现为理解相对正常,说话不流畅、话语不合语法。其损伤部位是额下回,位于大脑左前叶。(2)韦尔尼克失语症(Wernicke's aphasia),表现为说话流畅、话语基本合乎语法,但理解显然无能。其损伤部位是颞平面,位于左半脑的后部。布洛卡失语症患者的语法能力缺损,所以说话(造句)困难;但是,他们的话语理解还可以。这就引出一个问题:难道语言理解不需要语法?为了比较确切地了解这一点,就需要下列实验来证明。

实验一:句子—图片匹配测试(sentence-picture matching test)。例如:

(1) a. It was the $girl_i$ who_i[e_i chased the boy].

b. It was the boy_i $whom_i$[the girl chased e_i].

上面的a是主语分裂结构(subject-cleft construction),b是宾语分裂结构(object-cleft construction)。在被试(布洛卡失语症患者)听了句子后,让他们选择相应的图片。对于(1a)他们能做得很好,说明他们能很好地理解这种句子;但对于(1b)则做得很糟,说明他们不能理解这种句子。为什么?原来,在处理(1a)这种句子时,他们利用了默认施事在前的策略(agent-first default strategy)。他们用这种非语言的认知策略来作猜测,并且还每每得手。碰到(1b)这种句子,那种策略就失效了;因为,理解这种句子必须利用空语类跟其先行语的照应关系这种句法知识,但是布洛卡失语症患者的句法知识受损,就无法利用这种知识。据此,可以断定:该损伤部位是处理空语类等句法问题的神经基础。同时,这个实验还说明:布氏患者对于关系代词提供的形态格(如例1a, b中的主格、宾格等)也不能利用。

在视觉探测识别实验中,韦尔尼克失语症患者能作出正确的判断,但布氏患者却不能。于是,改作下列实验再行试验。

实验二:听觉—视觉交叉模式词汇启动(cross-modal lexical priming)。例如:

(2) The man liked the $tailor_i$ with the British accent who_i

$[e_i]$ claimed to know the queen.

这是一个主语关系从句结构（subject-relative construction）。在耳机上放这个句子，到空语类$[e_i]$处或其他地方，在屏幕上显示跟空语类的先行词 tailor 相关的探测词 cloth、或无关的控制探测词 weight（跟 accent 有关）作对照。要求被试大声读出，然后计算时间。结果，(i) 正常人（控制组）能正确地对空位填补（gap-filling）作出反应，并正确地理解全句的意义；(ii) 韦氏患者能正确地对空位填补作出反应，但不能正确地理解全句的意义；(iii) 布氏患者不能正确地对空位填补作出反应，但能大致理解全句的意义。换成下例再作试验：

(3) The priest enjoyed the $drink_i$ that the caterer was^1 $serving^2$ $[e_i]$ to the guest.

用探测词 wine 和 boat 在 1、2 处测试，结果大致跟例(2)一样。对此的解释是：韦氏患者能找到空语类跟先行词的句法依存关系，但不能建立起由动词决定的论元结构，即缺少给名词性成分指派语义角色的能力；因此，仍然不能理解句子。也正是由于缺少这种语义能力，因而他们造出来的句子在内容上是不合理、甚至是荒谬的。布氏患者不能找到空语类跟先行词的句法依存关系，于是对依赖于这种句法关系的论元结构的理解很困难，只得借助于施事在前这种非语法的认知策略。一旦碰到包含在宾语位置上有空语类的关系从句的句子（如 1b），这种认知策略就不再奏效，最终导致理解失败。这证明语言理解也必须有语法知识作支持。

从上面的实验可以得出这样一种可能的结论：(i) 布氏患者被损的大脑额下回的神经组织掌管句法依存关系等抽象的句法知识，也许还有短时记忆等职能；(ii) 韦氏患者被损的颞平面的神经组织掌管句法成分之间的语义关系等语义知识。从中得出的理论蕴涵是：(i)语言知识是分成句法、语义等模块的（modular）；(ii) 每一种类型的语言知识（句法知识、语义知识）在语言处理中有其特定的作用；并且，(iii) 它们在大脑中有特定的部位和相应的神经基础。

显然，这种类型的研究可以缩短认知科学和神经科学的距离，使认知研究这种主要依赖于各种假设的黑箱模型向神经生理学这种基

于实验的白箱模型过渡；最终，有希望形成一种研究人类思维的灰箱模型——一种半透明的工作范式。

4 认知科学和计算机理解自然语言

因为高级水平上的认知活动是一种串行的(serial)信息加工过程，可以理解为是一种在知识表示上的符号表达式的受规则约束的变换(即逻辑运算)，最终又可以还原为一定的算法和计算行为。所以，人的心智过程可以理解为符号处理的计算过程，人类的语言理解过程也可以理解为是一种在知识表示上的计算过程，这使得计算机理解自然语言在技术上具有可能性。[①] 因此，对语言的认知研究的自然延伸便是对语言的计算分析。于是，显然地认知科学对计算语言学应该有极为重要的认识论和方法论意义。

4.1 从认知研究走向计算分析

上文说过，认知科学有这样一个基本的信念：可以把计算机作为人类思维的模型，也可以用计算机来模拟人类的认知过程。由于语言是人类认知的最重要和系统的一个方面，因而人们自然会尝试用计算机来模拟人类的语言理解过程；从而造就了计算机科学中一个重要的研究领域——自然语言理解(natural language understanding)，并逐渐发展成一个综合性的前沿学科——计算语言学(computational linguistics)。那么，怎么才能让计算机理解自然语言呢？经典的人工智能方法是：首先把语言处理看作是一种问题求解过程，弄清人类在进行语言理解时的工作机制；然后把解题过程作出形式化的描述，再用一种形式化体系(formalism)来重写；最后用程序语言来表示，并在计算机上实现。

一般来说，这种类型的计算语言学研究分为如下三个步骤：[②]

第一步，数学建模。把需要研究的问题在语言学上加以形式化

① 详细的论证请看袁毓林(1996)。

② 参考冯志伟(1992)，第84页；钱锋(1990)，第26—27页。

(linguistic formalism),使之能以一定的数学形式、严密而规整地表示出来。也就是说,为有关的语言问题建立数学模型。包括选择恰当的形式语法(formal grammar)使得句子的结构能够用某种数学形式明确而清晰地表示出来,研究在这种形式语法之下如何分析句子构造的方法和步骤;选择恰当的语义表示体系使得句子的意义能够用某种数学形式明确而清晰地表示出来,研究在这种形式体系之下如何分析和表示句子的语义结构。

第二步,算法设计。把这种严密而规整的数学形式表示为算法(algorithm),使之在计算上形式化(computational formalism)。这就必须研究句子分析的严格的手续(procedures),并抽象成机械的、明确的、一步步逼近分析结果的步骤。

第三步,程序实现。根据算法用某种程序语言编写计算机程序,使之在计算机上加以实现(computer implementation)。

比如,Winograd (1983)可以说是认知主义计算语言学的杰出典范。他由下列两个问题激发灵感,尝试建立一种语言研究的认知范式(cognitive paradigm):

i. 一个人要说话和理解语言,必须具有哪些知识?

ii. 为了在交际中使用这些知识,人的心智是怎样组织的?

他把语言使用看作是一种以知识为基础的交际过程,认为人无论是说话还是听话都必须具有一定的知识;比如,词序规则、词汇和词的结构、语义特征、所指关系、时制系统、话语结构、说话人的态度、韵律规约、风格规约、世界知识等。在理论方面,他企图探讨人是怎样习得、运用这些知识的;在实际运用方面,他尝试用计算机来模拟人习得、储存、运用这些知识的过程,所以他又称这种范式为计算的范式(computational paradigm)。[①]

从信息加工过程的观点看,人说出一句话和理解一句话时,在大脑中有一个关于所描述的外部世界中的事物或事件的心理映象,可以称之为内部语言;而人处理语言的过程就是把外部语言转化为内

① 详见 Winograd (1983),pp. 1—34. 另外,参考黄奕(1985)对该书的介绍和评论。

部语言,经过加工后再由内部语言转化为外部语言的过程。计算机也可以用类似的过程来处理自然语言:首先确定一种语言的内部表示,然后寻求一种把所限定的语言子集中的语句转换为内部表示的方法。于是,让计算机理解语言的关键是:应能对一般的自然语言的句子作出语义解释,即设计一种一般的内部表示。内部表示是自然语言处理的关键,它影响着系统对语言知识和世界知识的描述和利用,因此也影响着整个处理系统。[①]

不同的学者由于对人类处理语言的心理过程的认识不同,因而采用了不同的理论和方法来建造自然语言处理系统。其中,一类系统比较重视句法分析,尽管所依据的语法理论各不相同。比如,Winograd(1972)年研制了关于积木世界的SHRDLU系统;该系统可以接受命令(通过一只机械手)对积木进行操作,回答有关积木世界所处的状态的问题。他认为句法需要解决的问题是:语言究竟是怎样组织起来表达语义的?他采用Halliday的系统语法(Systemic Grammar),把句法结构看作是生成句子的过程中一系列句法结构选择的结果。语义根据一定的外部世界模型作出推论来指示句法分析,从而得出句子的正确的语义解释。例如,在"I rode down the street in a car."中,只有运用世界知识(街道不可能在汽车里)作出推论,才能排除in a car作street的修饰语。Woods(1972)年设计了关于月球化学成分的LUNAR系统,该系统的句法部分根据Chomsky(1965)年的转换生成语法模型,分析出标准理论所指定的深层结构,再输入语义部分。语义部分根据句法上的深层结构再进行语义信息的分析。数据检索部分再根据输入句的语义编译成一种面向系统的形式语言(即查询语句),以便直接查询数据库,并最终产生结果(即回答)。Simmon(1973)年根据Fillmore的格语法(Case Grammar)建立了语义网络理论。另一类系统不作详细的句法分析,直接从语句中抽取语义信息。比如,Wilks认为,整段言谈的内容是由一些简单的基本信息构成的。一个复杂的句子也是由基本信息通过概念连结成实时的线性序列,而不是语言学家所认为的具有层次的树

① 详见杨抒(1988),第21—23页。

形结构。Wilks 于 1973 年用人工智能的方法设计了一个英法机器翻译的模型。这个模型不作句法分析，而是用一套“语义模板”来接受输入句中的信息。也就是说，该系统把源语言的输入语句直接处理为一种语义结构，作为一种中介成分，再据此生成目标语言的语句，也可以在这种中介成分上作谓词演算用于特定领域。Schank 认为人脑中存在着某种概念基础（conceptual base），语言理解的过程就是把语句映射到概念基础上去的过程。概念基础具有完善的结构，人往往能根据初始的输入预期可能的后续信息。句法分析对语言理解的用处不大，因为语言理解需要的是输入句的意思，而不是它的句法结构。计算机要理解语言，必须模拟人的心理过程；要像人一样根据上下文、环境、知识、记忆等作出预期（expectation），从而获取语义。句法只起一种指引的作用，即根据某些输入词语形成概念结构，预期它的句法形式，便于查找核实。Schank（1973）年提出了概念从属（Conceptual Dependency，CD）理论，建立了 MARIE 模型。上述这些不同的理论和方法，都是基于研究者对于“人是怎样理解语言的”这一问题的不同见解而发展出来的。也就是说，他们分别用不同的计算范式来实现其认知范式。[①]

4.2 两种计算范式：基于规则和基于统计

上面介绍的计算语言学的研究范式的特点是基于规则，即以知识（表示成规则）为基础的方法，通常称为人工智能的方法。这种方法假定：如果计算机要处理自然语言，那么它必须跟人一样具有句法、语义、语用、话语篇章、主题事物、周围世界等方面的知识和逻辑推理能力。因为人处理语言时的心理状态和心理过程就是这样的，计算机必须具有跟人相同和相近的知识才能处理自然语言。

而比较晚起的语料库语言学（corpus linguistics）采用的则是以语料统计为基础的方法，即基于概率的方法。这种方法认为：计算机并不能像人一样利用知识去理解语言，人们也无法把理解语言所需的各种知识形式化地表示成规则。有鉴于此，这种方法假定：如

① 详见杨抒（1988），第 22—26 页；范继淹、徐志敏（1980），第 9—19 页。

果我们能对数量很大的语言数据作出定量化的统计分析，那么我们就能对语言成分的分布和语言成分之间的关系等进行概率性的预测，从而补偿计算机缺乏知识和推理能力的缺点。[①]

虽然语料库语言学在词类标注等不需要涉及结构和语义的方面取得了诱人的成绩，但是在代词照应等涉及复杂的结构和意义的方面一时还难见功效。尽管从工程的角度，语料库语言学具有广阔的应用前景。不过，我们更偏爱基于规则的方法。因为，这种方法用 Hans Karlgreen 教授的话来说，就是“用计算的方法来制定人类语言行为的模型，并以此去了解人们怎样听说读写、怎样学习新知识和更新旧知识，又是怎样理解、存储和组织语言信息的”。他甚至认为，计算语言学的一个最根本的问题就是了解“人类的大部分活动在什么程度上能够简化成机械的操作”。[②] 显然，这种路子的研究对认知科学和语言学研究有更多的启发作用。

4.3 汉语的计算结构和计算模型

上世纪七十年代末，中国科学院心理研究所的李家治等先生进行计算机理解汉语的研究。他们用 Qillian 的语义记忆网络理论，开发了一个自动理解汉语的心理学模型。[③] 同时，中国社会科学院语言研究所的范继淹等先生进行人机对话研究，开发了一个铁路客运自动问答系统。这属于真正的语言学模型。为此，范先生对汉语的是非问句进行了非常系统的研究，并对语言与信息的关系、语法分析的理论和方法进行了全面的检讨和反思，提出了一种“语义短语语法”，对汉语语言学的研究具有很大的启发意义。[④] 七十年代中期，中国社会科学院语言研究所的刘倬等先生进行英汉机器翻译研究，致力于发展一种便于英汉对应的“中介成分”。其中，触及一些汉语句法、语义的深层次问题，对汉语语法研究也有一定的参考价值。[⑤]

---

① 参考桂诗春、宁春岩(1997)，第 138—149 页。

② 详见黄建烁(1991)，第 31 页。

③ 详见李家治、郭荣江、陈永明(1982)。

④ 详见范继淹、徐志敏(1981、1982)和范继淹(1986)中的有关文章。

⑤ 详见刘倬(1981)。

从80年代后期到90年代，北京大学计算机系/中国科学院计算机研究所的白硕先生进行了一系列基于语言学理论和方法的计算语言学研究。白硕(1995)指出：计算语言学旨在以自然语言处理(包括理解、生成、人机对话、机器翻译以及语音/文字输入的后处理等)为技术背景，揭示自然语言的词法、句法、语义、语用诸平面及其相互作用的计算结构，把语言学知识重塑成可以转化为产品的计算模型(第2页)。该书致力于研究语言学规则这种特殊形式的知识的发现的逻辑实质，全面地展示跟语言学知识发现有关的各个层次上的形式化机制——从数学建模、逻辑分析、算法描述、具体实现直到结果的语言学解释。作者采用语言学中经典的分布分析的思想，并针对真实语料的各种特点，结合汉语的实际，从数学、逻辑、算法和实现各个角度，全面阐述了从语料中发现确定性语言学知识(主要是词类和句法规则)的理论和方法。作者首先从数学角度讨论了分布理论的完善和推广，分别在词、短语、词结的划类问题上引入分布分析方法。作者在讨论词类及其划分的数学理论时，提出了词类划分的不动点理论、指出分布分析的任务是求解最大不动点，澄清了语言学界有关分布分析中含有"逻辑循环"的误解、证明了最大不动点在极限意义下的可计算本性、明确了分布分析方法的两个基本的逻辑前提：词的同一性和语言边界的明确性，从而解决在词类问题上"发现什么"和"能否发现"两大问题。在讨论发现句法规则的数学理论时，作者用构造性的方法建立一个基于句型推衍的变换规则系统，用以说明什么是基本句型和怎样从一些句型得到另外一些句型；其中，推衍规则包括句型推衍规则和环境推衍规则，它们都是重写规则(rewrite rules)；并阐明这种规则发现系统跟分布分析的关系：同分布关系和作为重写规则的推衍规则在本质上都是一种"替换"。就这样，作者从词的分布分析推广到了短语结构的分布分析，接下来他又把分布分析推广到词结(word complex，即超距相关的实词多元组，long-distance dependent word n-tupple，如："英语我十年前就会说了"中的"英语……说")。作者发现如果两个词结是同分布的，那么它们一定同时满足或不满足任何一个变换；所以变换是实词多元组和多元句法环境之间的一种推衍关系，词结是变换下的不变量、是多元环境

的填充物，而多元环境则是某一句法结构中抠掉了词结的剩余部分；由于词结是以各种不同的多元环境作为分布框架的，因而变换分析就是词结的分布分析，通过变换分析可以给词结进行分类。这样，句子可以看作是由词结加上环境构成的，句子语义恰好可以分解为词结的语义加上环境的语义。比如：

"河不过了"，指的是撤销"过河"的意愿；

"饭不吃了"，指的是撤销"吃饭"的意愿；

多元环境"不……了"的语义为"实现事件 E 的愿望撤销了"，加上由词结"过……河、吃……饭"的意义正好是句子的意义。作者甚至希望通过词结的分布分析，来归纳词结中的从属成分的语义格。其根据是词结的同分布类跟内部语义角色关系和外部组合能力相同的语义结构类是大致对应的，这样，同分布的词结的相同位置上的从属成分的语义格是相同的，比如，上例中"河、饭"的语义格是一致的。这在方法论上，对语言学研究无疑是有很大的启示作用的。

4.4 基于认知并面向计算的汉语语法研究

上文介绍的那种以人类认知为基础的计算语言学研究，催生了一种基于认知并面向计算的语言研究路子（a cognition-based and computation-oriented approach of linguistic study）。这种研究路子在汉语语法的研究方面已经进行了一些实践，并收到了一定的成效。比如，袁毓林（1993）§5 指出：人类的语言理解除了需要句法、语义等语言学知识之外，还依赖于常识。例如：

(1) He hit the car with the rock.（他用石块砸车子）

(2) He hit the car with the dented fender.（他砸装有前挡板的车子）

人们凭借他们对于 hit 与 rock（动作—工具）、car 与 dented fender（整体—部分）之间的关系这种世界知识（world knowledge），来决定这两句的语法构造（with the rock 作状语修饰 hit the car、with the dented fender 作定语修饰 car），最终得出正确的语义解释。但是，像 hit 与 rock 的"动作—工具"关系、car 与 dented fender 的"整体—

部分”关系之类的常识很难穷尽，也不易于形式化。为此，作者提出了一种新的思路：把部分跟语言理解相关的常识化解为一种句法、语义知识，通过语言学的句法、语义刻画手段来形式化；其途径之一是通过名词的配价研究，把关于事物之间的各种复杂关系的常识转化为一种代表事物的有关名词之间的句法、语义关系。在这种思想的指导下，袁毓林(1992、1994)分别研究了现代汉语中的一价名词和二价名词的句法、语义特点，并结合认知科学的研究成果，用扩散性激活的语义记忆机制和非单调推理的逻辑机制，来分析有关句子的语义解释问题。例如：

(3) 这种酒很淡。(a. 味儿淡 > b. 颜色淡)

(4) 这种花很淡。(a. 颜色淡 > b. 味儿淡)

这种语义理解上的不平行性只能从语义记忆和语义推导的方式上寻求解释。比如，名词“酒”可以激活〔液体、饮料、刺激性的味道、颜色……〕等一组语义，名词“花”可以激活〔植物的器官、观赏性的颜色、味道……〕等一组语义，形容词“淡”可以激活〔(味道、颜色)不浓、(含量)稀薄、(态度)不热情……〕等一组语义。人们根据常识推断，酒作为一种有特别味道的饮料，〔味道〕是它的强特征，就直接把“酒淡”理解为“酒的味儿淡”。因为，根据缺省推理(reasoning by default)的原理“除非特别说明，可以默认某个命题总是成立的”，听话人有理由相信：如果说话人想表达“这种酒颜色很淡”，那么他一定会把表示酒的弱特征的“颜色”说出来。同样，“这种花很淡”中花的强特征完全可以省略，在语义解释时必须优先补入。有意思的是，白硕在九十年代后期，尝试用范畴语法的演算规则来建立一个语言理解系统，为网上的信息快速查找服务。在这个系统中，他除了利用动词、形容词的配价信息外，大量地把名词配价研究的成果吸收了进去，增强了该系统的表示能力和推演能力。

袁毓林(1996)甚至希望用扩散性激活的语义记忆模型和缺省推理的非单调逻辑来建立一种语言理解的微观机制，用以解释同一句子中不同词项之间的语义连结和制约关系；并以此来揭示人脑处理语言信息的某种心理过程，从而为认知心理学和计算机理解自然语

言提供强有力的语言学支持。作为案例，作者着重分析了下列例子：

(5) a. 这房子很大 ├ b. 这房子面积很大

(6) a. 这箱子很大 ├ b. 这箱子体积很大

(5a)和(6a)的句法、语义构造是一样的，但是语义解释却很不一样。对此，可以从认知的角度假设：(1) 大脑中语义储存的方式是网络(network)式的，语义提取的方式是扩散性激活(spreading activation)式的。并且，由于常识和生活经验(房子用以住人、箱子用以装物)的作用，人们在听/看到"房子"这个词时，〔面积〕这一语义节点优先激活，它跟其他词的语义节点的连接权值增大；人们在听/看到"箱子"这个词时，〔体积〕这一语义节点优先激活，它跟其他词的语义节点的连接权值增大。(2) 语义推导的方式是基于知识的缺省推理。虽然"大"的语义可以跟〔面积、体积、数量、强度、力量〕等语义节点相连接，但是人们在听/看到"房子大"时可以直接理解为〔房子的面积大〕，听/看到"箱子大"时可以直接理解为〔箱子的体积大〕。因为，听话人相信说话人一定遵守交际的缺省约定，如果说话人要表达〔房子的体积大〕或〔箱子的面积大〕这种意思，那么他必须特别声明，不能省去"体积"或"面积"这类词语。

非常有意义的是，姬东鸿、黄昌宁(1996)在建立关于汉语形容词跟名词的语义组合的计算模型时，还真的运用袁毓林(1994)提出的语义扩散性激活和缺省推理的机制、语义特征强弱的优先顺序以及相关的规则和策略，作为消解由多重属性继承引起的冲突的机制。例如：

(7) 王明很难受。(a. 心里难受 > b. 肚子难受)

(8) 这孩子很灵。(a. 脑子灵 > b. 耳朵灵)

(9) 衣服很大方。(a. 样子大方 > b. 领子大方)

这里名词"王明"既具有心理属性、也具有生理属性，而形容词"难受"既可以描写心理属性、也可以描写生理属性。这样，当名词的语义跟形容词的语义相互组合时，就势必会发生多重属性的冲突的问题。怎么来消解这种冲突呢？根据袁毓林(1994) §5 提出的心理属性强于生理属性、整体属性强于局部属性的优先顺序，可以用这种属性继

承的优先规则来解决这一问题。像这种基于认知的语言研究，计算语言学研究者和心理语言学研究者都是比较感兴趣的。

袁毓林(2004)则尝试用认知图式(cognitive scheme)的概念来分析词的意义和用法，并从中引导出可以转换成算法化的规则的形式表示。例如：

(10) 满身是汗 ～ 全身是汗　满商场的人 ～ 全商场的人

(11) 满脸是汗 ～ *全脸是汗 *满公司的人 ～ 全公司的人

对于“满”和“全”在意义和用法上的不对称性，可以用隐喻投射(metaphor projection)理论来解释：跟“满”相关的语言表达以容器(container)隐喻为基础，跟“全”相关的语言表达以套件(suite)隐喻为基础。在人们的观念中，身体和商场既可以看作是容器、又可以看作是套件；但是，“人”跟“公司”这种抽象的机构难以形成容物跟容器的关系，“脸”这种人体部件一般不再分解为几个更小的部件，即它不是套件。值得注意的是，在以容器隐喻为基础的语言表达中，容器在空间上具有拓扑可变性(立体、平面等)：

(12) 满杯子啤酒～满头白发～满纸荒唐言～满枝头麻雀～满门抄斩～满眼春色

说明容器隐喻等在心理上的表征应该是抽象的图式，是一种意象图式(imagery scheme)。不同的隐喻反映人们感知事物和事件时的不同的认知方式，从而构成了不同的意象。意象可以抽象为结构化的图式，图式可以分解为结构成分及其构成方式。如果找出隐喻表达的构成成分及其结构关系跟相应图式的构成成分及其结构方式之间的映射关系，就可以用产生式写出算法化的关于隐喻表达的语义解释规则。比如，对于容器隐喻来说，其意象图式的结构成分是一个边界，它把相关的空间划分为内部和外部两个部分，从而在人的心理上形成一个容器的构型。抓住了这一点，我们就可以给出从容器隐喻表达的句法形式到语义表达的形式化的、并且经过调整后可以是算法化的规则系统。假如把“满桌子糖果、满桌子的糖果、满桌子是糖果”等格式合记作 S1：满＋NP_1＋(的/是＋)NP_2，那么可以用一阶谓词逻辑写出 S1 的如下语义解释规则 R1：

if：满＋NP_1＋（的/是＋）NP_2；then：

{i. 'NP_1' is-a CONTAINER，'NP_2' is-a CONTENTS；'NP_2' is-in 'NP_1'；

ii. $\exists_y, \forall_x[\text{is-in}(x, y)] \rightarrow x =$ 'NP_2'，$y =$ 'NP_1'；

iii. CONTAINER has many SUB-SPACE，i. e.，$y = y_1 + y_2 + \cdots + y_n$；

iv. CONTENTS has many SUB-CONTENTS，i. e.，$x = x_1 + x_2 + \cdots + x_n$；

v. $\forall_{y_i}, \exists_{x_i}[\text{has}(y_i, x_i)] \rightarrow x_i \in$ 'NP_2'，$y_i \in$ 'NP_1'，i＝1，2，…，n}

如果把语句实例"满桌子（的/是）糖果"代入 R1，那么可以得出如下的语义表达式 M1：

'桌子'是容器，'糖果'是容物；'糖果'在'桌子'上；

存在着一张桌子，所有的'糖果'都在这张'桌子'上；

'桌子（面）'有许多子空间，'糖果'有许多子集；

'桌子（面）'的每一个子空间中都有一些'糖果'。

对于套件隐喻来说，其意象图式的结构成分是一个整体和若干个部分、一个体现各部分如何构成整体的构型。抓住了这一点，就可以参照上文对容器表达的计算分析，把套件的各部分看作是一个个容器，于是套件就成为一套容器；相应地，在这些容器中的容物也成为一套离散的容物。这样，就可以给出从套件隐喻表达的句法形式到语义表达的形式化的、并且经过调整后可以是算法化的规则系统。假如把"全身伤痕、全身的伤痕、全身是伤痕"等格式合记作 S2：全＋NP_1＋（的/是＋）NP_2，那么可以用一阶谓词逻辑写出 S2 的如下语义解释规则 R2：

if：全＋NP_1＋（的/是＋）NP_2；then：

{i. 'NP1' is-a-set-of CONTAINERS，'NP_2' is-a-set-of CONTENTS；

'NP_2' is-in 'NP_1'；

ii. $\exists y, \forall x$ [is-in (x, y)]$\rightarrow x=$'NP_2', $y=$'NP_1';

iii. CONTAINERS is-a SET consists of many SUB-SET, i. e. , $y=y_1+y_2+\cdots+y_n$;

iv. CONTENTS is-a SET consists of many SUB-SET, i. e. , $x=x_1+x_2+\cdots+x_n$;

v. $\forall y_i, \exists x_i$ [has (y_i, x_i)]$\rightarrow x_i \in$'NP_2', $y_i \in$'NP_1', $i=1,2,\cdots,n$;

vi. $\lambda(x_1, x_2, \cdots, x_n)$ [is-in (x_1, y_1) & is-in (x_2, y_2) &… & is-in (x_n, y_n)];

vii. $\Sigma_x = x_1+x_2+\cdots+x_n$}

如果把语句实例"全单位(的)职工"代入 R2,那么可以得出如下的语义表达式 M2:

'单位'是一套容器,'职工'是一批容物;'职工'在'单位'中;

存在着一个'单位',所有的'职工'都在这个'单位'中;

'单位'有许多子集(即部门),'职工'有许多子集,

'单位'的每一个子集(即部门)中都有一个'职工'的子集;

每一个子单位(即部门)中的职工子集的总和就是'全单位(的)职工'。

这种研究的目标是,从隐喻的角度分析诸如此类的词语同现限制问题,并把隐喻分析提升到意象图式的抽象水平。藉此,希望把语言的认知解释转换成算法规则和形式表示,从而实现认知和计算的统一。

5 结语:入门的台阶

许多年轻的朋友问:要从事计算语言学方面的学习和研究,应该有哪些知识上的准备?这可以从计算语言学的定义上说起。粗略地说,计算语言学是一门用计算机并为计算机研究语言的综合性学科。用计算机来研究语言,不仅指把计算机这种电子装置作为语言研究的辅助工具,比如,用计算机收集语料、分类整理、分布统计、提

取各种数据等。这跟化学、物理学、生物学中的计算化学、计算物理学、计算生物学有点相近,它们或者运用简单的方程和算法在计算机上进行大量的重复运算,或者用计算机对实验结果进行十分精细的计算分析、反复提高以得到一种新的理论。更重要的是指用计算机科学的理论、概念和方法来研究语言,我们认为这一点才是计算语言学更本质、更深刻的特点。比如,白硕(1995)用理论计算机科学的观点剖析当代语言学的方法、并进行计算模拟的做法,在一定程度上展示了这类研究的理论魅力和实用价值。

在这方面,计算神经科学(computational neuroscience)为我们提供了一个光辉的典范。作为神经科学的一个新的分支,计算神经科学通过建立脑模型来阐明神经系统信息加工的计算原理,以了解人和动物的神经系统是怎样使用它的微观组件及其相互作用来表征和处理信息的。具体的做法是:把神经科学对脑结构和机能从整体、细胞和分子水平上进行的生物学研究作出数学概括、找出规律和算法,并运用现代数字计算机或人工神经网络加以模拟;其最终目标是:揭露脑的电信号和化学信号,寻求如何表达和处理神经信息、并在智能活动中发生变化的规律。这种脑模拟研究通常使用简化的脑模型(simplifying brain models)。因为,即使是最成功的生物脑模型也不能揭示脑组织的全部实际功能;所以,计算神经科学需要抓住重要的原理进行简化模拟。简化模型的研究必须提供建立模型的理论框架、算法及其约束条件,而这种简化模型中的算法及其约束条件往往可以通过现代数字计算机或神经计算机来加以实现。可见,计算神经科学并不意味着大量的计算、也不意味着一定要使用现代计算机,而是要对大脑的认知过程进行表征,把其信息加工过程和信息存储过程跟计算机进行类比,从中得到新的概念和数学表达。比如,Hopfield 模型的建立并没有借助计算机进行大量的数值计算,但是这种模型有助于对大脑获取信息(即学习)和提取信息(即记忆)过程的理解;因此,这种数学模拟仍是计算神经科学的一个组成部分。同样,我们认为,计算语言学并不意味着大量的计算、也不意味着一定要使用现代计算机,而是要对大脑中的语言处理过程进行表征,把语言信息的加工、存储过程跟计算机进行类比,从中得到新的概念和数

学表达，以形成便于机器处理的语法规则或语法形式体系。计算神经科学致力于寻求理解智能活动的神经基础的新概念、新算法，并在把新算法及其约束条件跟当代各类计算机进行类比中，发现设计智能化计算机、智能化机器人和智能化武器的新原理。并且，计算神经科学提出的脑模型能够对神经系统的某些行为作出可以验证的预测，从而较早地预见到生物脑研究工作的成果。因此，计算神经科学对大脑的模拟研究，不仅为信息科学的发展提供了坚实的神经科学基础，而且对神经科学和心理科学的发展也起着巨大的推进作用。[①] 我们则希望，采用理论计算机科学的观点所进行的计算语言学研究，不仅对信息科学、神经科学和心理科学起推动作用，而且对语言科学的发展起巨大的推动作用。

为计算机研究语言，指为了计算机能处理自然语言而研究语言。这包括两方面的工作：(1) 对自然语言的结构和意义规律进行挖掘，提炼出便于形式化和算法化的句法、语义规则，建立合适的语法学理论模型，来更好地组织语言的句法、语义规则；(2) 把语言学家对语言的句法、语义、语用诸平面上的研究成果进行数学概括，用某种形式化体系来组织和表示语言的结构和意义规则，再找出恰当的算法来描述句子的结构分析或语义解释的严格的步骤(procedure)，最后根据算法用相应的计算机语言来编程实现。上面(1)所说的工作本应完全由理论语言学家来承担，但是，由于理论语言学关心的方面不一定跟计算语言学家一致，因而计算语言学家常常会发现：语言学中并无他们想要的句法、语义规则或语法理论模型；于是，计算语言学家只得亲自动手来寻找句法、语义规则，甚至建构更适合计算机的语法理论模型。在为计算机研究语言这一点上，计算语言学有别于计算化学和计算神经科学。在计算化学中，并没有为计算机研究化学这种任务；在计算神经科学中，也没有为计算机研究神经的结构和功能这种任务。那么，为什么计算语言学要特别地强调为计算机研究语言这一点呢？原因可能有两点：(i) 语言学的研究对象是自然语言，语言学的研究工具(用以描写语言现象、表述语言规律、总

① 关于计算神经科学，参考沈政、林庶之(1992)，第44—49页。

结研究结果)也是自然语言。也就是说,自然语言既是语言研究的对象语言(object language),也是语言研究的元语言(metalanguage)。由于计算机无法直接理解自然语言,因而首先必须把用自然语言表述的语言规律形式化、符号化。(ii) 语言是一种心智(mind)现象,是跟人的认知、心理密切相关的;为了让计算机能理解自然语言,必须以计算机为信息加工模型来考察人类语言理解的心理过程,以便在计算机上模拟实现。

有了这样一番理解,那么显而易见,要从事计算语言学方面的学习和研究,首先应该了解或掌握语言学和计算机科学方面的一些基础知识。语言学方面,包括语音学、实验语音学、音系学、句法学、语义学、语用学、话语语言学等;计算机科学方面,包括体系结构、数据结构、算法理论、程序语言、形式语言和自动机理论、复杂性和可计算性理论、人工智能原理等。再有是这两门学科的综合学科——计算语言学。此外,心理学方面,包括认知心理学、神经心理学、实验心理学、语言心理学等。还有,数理逻辑方面的知识也是不可缺少的;再奢侈一点,脑科学、神经生物学、西方现代哲学(特别是心智哲学、科学哲学)也是应该关注的。不过,上面涉及的这么多的门类和内容,不一定非得在短短几年内全部都来学一遍,而是要求放宽眼界,有长远的目标和计划。一般来说,在六到十年时间内,程度深浅不同地摸一遍,这应该是大家都可以做到的。

现在说说怎么着手进行这方面的研究。在很大程度上,这要取决于每个人的不同的环境条件。比如,我的师兄陈小荷,在北大攻读博士学位期间,并没有接触多少计算语言学方面的知识,博士论文做的是江西丰城话的语法;毕业以后分配到北京语言学院语言信息研究所,工作的需要促使他不断地学习这方面的知识,学习编程、尝试建语料库,参加905语义工程,促使他思考和研究面向工程的语义分析体系问题,又从自动句法分析的角度考虑汉语词类问题,等等,逐步进入这一领域。王惠在北大中文系读书时做的硕士论文是《从及物性系统看现代汉语句式》,毕业后分配到北京大学计算语言学研究所,工作的需要,促使她逐步了解中文信息处理方面的知识,并着力于对面向中文信息处理的语法信息词典和语义词典的研究。詹卫东

在浙江大学中文系读本科时已经接触了语言信息处理方面的有关内容，到北大中文系跟陆俭明先生学习语法学，同时在计算语言学研究所接受俞士汶先生指导，博士生期间继续这种模式，在计算语言学方面有比较好的基础和训练；毕业后留在中文系，同时在计算语言学研究所承担研究任务，先后对面向中文信息处理的汉语短语结构的约束条件和语义知识的表示等问题，进行了比较系统和深入的考察，形成了一些独到的见解。但是，大多数人可能会跟我一样，在中文系读书，毕业后又在中文系教书。因此，说说我的学习经历，也许对大家也有一定的借鉴作用。上世纪 80 年代初，我在《百科知识》上看到理论计算机科学、人工智能、语言信息处理方面的文章，又在《国外语言学》和《中国语文》上看到计算语言学、特别是跟汉语相关的人机对话、机器翻译方面的文章，开始对计算语言学很神往。1984 年在杭州大学中文系上研究生，在《数理逻辑》课上，经常听到邱国权老师讲数理逻辑和人工智能、数理逻辑和机器处理语言的关系，在他的鼓励下学习了 Basic 编程语言。也看了一些计算机方面的书籍，特别对范继淹先生那种语法研究和信息处理互相结合、互相促进的研究模式顶礼膜拜。1985 年春天，不仅在杭大聆听了范先生关于人机对话的讲座，还在宾馆向范先生请教了汉语语法研究的门径问题，范先生的精彩指点，使我深受教益，有没齿难忘之感。1987 年到北大攻读博士学位，参加由中文系朱德熙和陆俭明等先生、计算机系马希文和林建祥等先生、心理学系王甦等先生、哲学系赵光武等先生组织的人工智能的哲学基础的讨论班，参加了计算机系青年教师王培组织的一个关于人工智能和认识论方面的讨论班，又参加了林建祥老师主持的机器学习讨论班，还经常跟马希文先生的博士生白硕一起讨论语言分析及其计算机处理问题。1990 年分配到清华大学中文系工作，在罗振生老师的奔走和帮助下，得以利用清华大学智能技术与系统国家实验室的机房，一边学 C 语言，一边在机器上学习编程序。在学习计算机科学技术方面的有关知识的同时，不断地参加中文系和计算机系的有关讨论和研究生的开题和答辩，还协助罗振生老师指导计算语言学方面的研究生。我基本上是站在语言学的角度，从理论上思考计算机理解自然语言问题。抓住跟语言理解有关的知识

的形式表示问题，探讨动词、名词配价的作用；同时，考虑跟语义推导有关的认知机制和逻辑机制问题。《语言的认知研究和计算分析》一书中的好几篇文章，就是这样的背景上形成的。现在回过头来，觉得每一个语法学者都可以做的工作是：挑某一种自己觉得是比较特别、也比较有趣（好玩）的语言现象，比如某种语法格式；想一想（内省）你自己是怎样从这一符号串上得出这种语法形式的意义的，要理解这种语法形式所表达的意义，需要哪些句法、语义等语言内的知识，还需要哪些百科知识类的常识，需要遵循什么样的规约、作出怎样的逻辑推导，等等。然后，对这一格式及其有关实例作出具体的描写和分析，努力找出使这一格式合格的句法、语义约束条件，并尽量明确地表示出来；然后再考虑哪些常识和推理方式参与了这一理解过程，它们是怎样跟有关的句法、语义知识发生交互作用的，能不能用一个比较抽象和统一的模型把这一语言理解过程（各种因素及其作用方式）表达出来。能做到这一步，也等于是为这一语法格式的意义理解建立了一个初级的逻辑模型。至于怎样精炼化为严格的数学模型、判定该形式模型是否具备可计算性、度量计算的复杂性，以及算法设计、程序实现等工作，完全可以由计算机专家来做。因为计算语言学的工作是一种系统工程，语言学家只要提出一个尽可能可靠、简单的初步模型就可以了。因此，对于语言学家来说，计算语言学工作主要任务是：尽可能详尽而明确地描写有关语言现象，探明有关因素的作用方式及其关系，揭示使这一语言现象成为合格、可接受的各种约束条件。至于你会不会编程序、懂不懂算法理论和数据结构，倒不一定太重要。当然，最好大家对计算机是怎样工作（特别是怎样处理自然语言）的原理有所了解；这样可以帮助我们了解什么样的约束条件和规则是重要的，什么样的语言学模型对信息处理是有用的。由于我们是语言学者，因而我们从认知、计算等角度思考语言问题时，目的仍主要在于检验各种语言理论和分析方法的效能，希冀以计算机为参照，来提高语言学的研究水平，使语言学真正成为一门严格意义上的科学。

参考文献

白　硕 (1995)《语言学知识的计算机辅助发现》,科学出版社。

白　硕 (1996)《语言实用主义》,罗振生、袁毓林主编(1996)《计算机时代的汉语和汉字研究》,清华大学出版社。

范继淹 (1986)《范继淹语言学论文集》,语文出版社。

范继淹、徐志敏 (1980)《自然语言理解的理论和方法》,《国外语言学》第 5 期。

范继淹、徐志敏 (1981)《关于汉语理解的若干句法、语义问题》,《中国语文》第 1 期。

范继淹、徐志敏 (1982)《RJD-80 型汉语人机对话系统的语法分析》,《中国语文》第 3 期。

冯志伟 (1992)《计算语言学对理论语言学的挑战》,《语言文字应用》第 1 期。

冯志伟 (1996)《自然语言的计算机处理》,上海外语教育出版社。

桂诗春、宁春岩 (1997)《语言学方法论》,外语教学与研究出版社。

郭承铭 (1993)《认知科学的兴起与语言学的发展》,《国外语言学》第 1 期。

黄建烁 (1991)《计算语言学研究综述》,《国际学术动态》第 4 期。

黄　奕 (1985)《认知过程的语言》,《国外语言学》第 3 期。

姬东鸿、黄昌宁 (1996)《汉语形容词和名词的语义组合模型》,*Communications of COLIPS*(中文与东方语言信息处理学会通讯), Vol. 6, No 1.

李家治 (1985)《国外认知科学介绍》,《思维科学》第 2 期。

李家治、郭荣江、陈永明 (1982)《机器理解汉语——实验 I》,《心理学报》第 1 期。

刘　倬 (1981)《JFY-II 型英汉机器翻译系统概述》,《中国语文》第 3、4 期。

钱　锋 (1990)《计算语言学引论》,学林出版社。

沈　政、林庶之 (1992)《脑模拟和神经计算机》,北京大学出版社。

石纯一、黄昌宁、王家廞 (1993)《人工智能原理》,清华大学出版社。

翁富良、王野翊 (1998)《计算语言学导论》,中国社会科学出版社。

杨　抒 (1988)《自然语言的认知模型》,《计算机科学》第 3 期。

袁毓林 (1992)《现代汉语名词的配价研究》,《中国社会科学》第 3 期。

袁毓林 (1993)《自然语言理解的语言学假设》,《中国社会科学》第 1 期。

袁毓林 (1994)《一价名词的认知研究》,《中国语文》第 4 期。

袁毓林 (1996)《语言的认知研究和计算分析》,《语言文字应用》第 1 期。

袁毓林 (2004)《容器隐喻、套件隐喻及相关的语法现象——词语同现限制的认知解释和计算分析》,《中国语文》第 3 期。

章士嵘 (1992)《认知科学导论》,人民出版社。

周光召（1995）《迈向科技大发展的新世纪》,《中国科学报》5 月 29 日。

Fodor, D. Janet (1995) Comprehending Sentence Structure, in Lira R. Gleitman and Mark Liberman (1995) *An Invitation to Cognitive Science*, Vol. I: *Language*, p. 209—46. The MIT Press.

Gardner (1985) *Mind's New Science*, Basic.

Gazdar, G. & Mellish, C. (1987) Computational Linguistics, in J. Lyons, etc. (ed.) *New Horizons in Linguistics* 2. Penguin Books.

Grishman, Ralph (1986) *Computational Linguistics: An Introduction*. Cambridge University Press.

Halle, Morris (1973) A Window into Man's Mind, in Eric p. Hamp (1973) (ed.) *Themes in Linguistics: 1970s*, Mouton.《洞察人类心智的窗口》,曹今予译,沈家煊等校,《国外语言学》1984 年第 1 期。

Halvorsen, Per-Kristian (1988) Computer Applications of Linguistic Theory, in F. J. Newmeyer (ed.) *Linguistics: The Cambridge Survey*, Vol. II, *Linguistic Theory: Extensions and Implications*. Cambridge University Press.

Lyons, (1968) *Introduction to Theoretical Linguistics*, Cambridge University Press.

Piattelli-Palmarini, Massimo (1980) (ed.) *Language and Learning: The Debate between Jean Piaget and Noam Chomsky*, Harvard University Press.

Winograd, Terry (1983) *Language as a Cognitive Process*. Addison-Wesley Publishing Company, Inc. 中文简介请看黄奕(1985)。

2003 年 12 月初稿,2004 年 6 月改定

（收入刘丹青主编《语言学前沿与汉语研究》,上海教育出版社,2005 年）

面向当代科技的语言研究的理论和方法

本文是一篇演讲稿，讲述搞语言研究的人怎样吸收当代科学技术的有关理论和方法上的成果，形成一些具有开拓性的理论意义和应用价值的课题。具体地说，怎样从认知心理学和计算机科学和技术的角度，来形成一些对解决汉语语法比较有效的理论和方法。主要通过自己做的四个研究案例（语言理解、语义演算、词类辨认、定语排序），来讨论上面提出的这个问题，从而展示认知语言学和计算语言学相互结合的一种可能的研究路子。

0 引 言

我今天要讲的题目是“面向当代科技的语言研究的理论和方法”。主要想谈的问题是搞语言研究的人怎样面对当代科学技术的发展，形成一些具有开拓性的课题。更具体一点说，就是怎样从认知心理学和计算机科学技术的角度来寻找一些对解决汉语语法比较有效的理论和方法。这里边涉及到一些比较关键的当代科学技术方面的概念，比如：认知、计算以及认知和计算的关系等。但是，我不准备直接从这几个概念上讲，因为那比较抽象，也不好懂。我打算从四个研究案例，即自己做过的几个工作上来回答上文提出的这个问题。

1 语言理解的心理机制和逻辑机制

先讲第一个问题，关于语言的意义理解的一个例子。

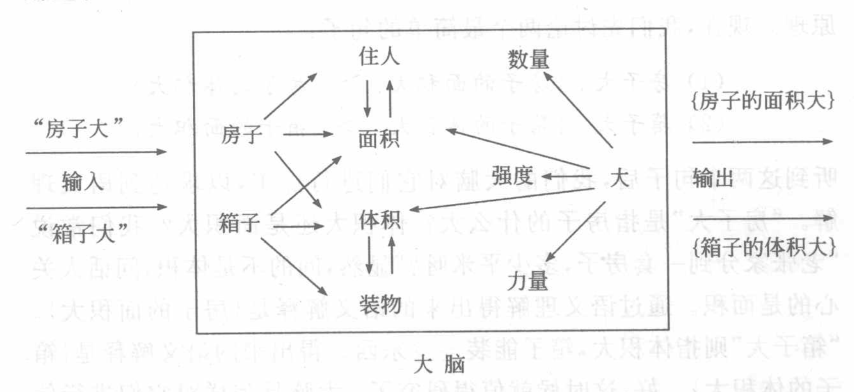
大家知道，人的大脑现在仍是一种密封的机构，无法打开。就是说，你无法打开一个活人的大脑去看看内部的工作机制是什么、它是怎样来解决问题的，所以只能从另外一个角度来研究大脑的各种生

物机构是怎样工作的。有一种办法叫做黑箱模型：虽然人的大脑是一个密封的黑箱，我们无法打开来看；但是，可以比较这个黑箱的输入和输出这两方面的情况，然后来建造、构拟大脑的工作机制是怎样怎样的。也就是说，通过这种输入和输出的关系，来假设大脑的工作原理。现在，我们先讨论两个最简单的句子：

(1) 房子大。{房子的面积大}＞{房子的体积大}
(2) 箱子大。{箱子的体积大}＞{箱子的面积大}

听到这两个句子后，我们的大脑对它们进行加工，以求达到语义理解。“房子大”是指房子的什么大？体积大还是面积大？我们常说“老张家分到一套房子，多少平米呀？”显然，问的不是体积，问话人关心的是面积。通过语义理解得出来的语义解释是{房子的面积大}。“箱子大”则指体积大，箱子能装多少东西。得出来的语义解释是{箱子的体积大}。好，这时候就值得研究了：大脑是怎样对它们进行信息加工，从而得出这两种不同的语义解释的。因为从常识上看，房子和箱子都是三维的，都是有面积和体积可言的；但是“房子大”得出来的语义解释是{房子的面积大}，而“箱子大”得出来的语义解释是{箱子的体积大}。搞心理学的人就要研究它，要追究人的脑子里面有哪些心理结构和认知结构，经过了怎样的认知加工过程，才能达到这么一种语义理解。作为一个搞人工智能的人，他关心的是这么一个语言理解过程，怎样用一套形式化的规则把它表示出来，建立一个可计算的数学模型，然后让机器去模拟，并通过程序在机器上实现这样的理解。如果说他的机器是一个具有智能的机器，对语言有很好的理解能力的话；那么给它输入“房子大”，输出的语义理解应该是{房子的面积大}，而不是{房子的体积大}。如果做不到这一点，说明这台机器还不够智能化，甚至是没有智能的。人脑里面到底有什么样的认知结构，通过什么样的认知过程来完成这么一个解决问题的任务？我们认为大脑中的认知结构是一个网络式的结构，也就是说，语义在大脑中的储存是以一种网状的方式。而不是像传统上那样，认为大脑中有一部心理词典，这种心理词典中储存一套词汇，这些词汇表现为一条一条离散的词项，它们以声音或意义等线索编排在一起。相

反，现在我们更相信大脑中的心智词典是一种由语义结点构成的网络。当你听到一个词时，它激活（activate）了网络上的一片相关的语义结点。比如，我们可以这样来表示“房子大”与“箱子大”的语义理解过程：



大 脑

一个人听到“房子”和“箱子”的时候，脑子里至少激活了［面积］和［体积］这样两个语义结点；同时“大”可以指体积大，也可以指面积大，还可以指数量大，甚至可以指力量大、能力大、重量大……。有了这样一个网络，能不能给出一种有效的计算方法，使我们听到“房子大”时，把［房子→面积→大］→｛房子的面积大｝这条语义路线接通；听到“箱子大”时，把［箱子→体积→大］→｛箱子的体积大｝这条语义路线接通。其实，我们听到“房子大”后，除了激活［面积］、［体积］这两个语义结点之外，还有另外一些语义结点，如房子是供［住人］用的，而箱子是供［装物］用的。而［住人］和［面积］存在着相互作用的关系，［体积］与［装物］之间也有相互作用的关系。正是通过语义结点之间的相互作用关系，把某些语义结点强化了，同时把另外一些语义结点抑制了。在这里，“房子”尽管激活的是［面积］和［体积］这两个语义结点，由于［住人］这个语义结点的加入，一下子就把［面积］这个语义结点的权值（weight）加重了，所以［面积］被突显成为强特征，［体积］被压抑成为弱特征。因此，当我们听到“房子”时，首先想到的是［面积］这个特征；而听到“箱子”的时候，由于［装物］这个特征的加入，一下子把［体积］这个语义结点的权值加重了。在“大”激活的几个语义

结点中,“房子”跟“大”组合的时候,就遵循一种基本的规约:主语名词的强特征优先与谓语形容词的选择特征进行匹配,其他特征被压抑和淘汰,最后输出的结果便是{房子的面积大}。在这种语义网络上,经过推理,得到这样一种语义解释。这样一种网络,就是我们所谓的“认知结构”,它的生物基础是一种神经网络系统。神经网络目前在人工智能研究上很热。传统的人工智能认为,我们处理问题是根据一条条规则的先后顺序来进行的。现代神经网络理论认为,如果是那样的话,一个人要做一件事情就根本来不及做。因为一条条规则的调动要按先后顺序,是需要相当的时间的;但实际上人们做事情的时候,是很快就做完了的,不可能有那么长的时间供你按部就班地去做。所以,在人工智能研究中,许多人尝试运用神经网络,来代替传统的人工智能中的串行规则。上面讲的是认知结构,下面我们接着讲认知过程。

人类的认知过程在许多情况下是一个推理的过程,其中经常用到的一种推理叫做“缺省推理”(reasoning by default)。这种推理形式生活中比比皆是,但传统的逻辑一般不研究它。因为它具有非单调性的(non-monotonous)特点,表现为一组前提及由其推出的一组结论,有可能是不一致的,会违反逻辑上的排中律。直到计算机科学上的人工智能兴起,才研究这种逻辑。通俗地说,这种基于缺省推理的非单调逻辑的要义是:在一般的日常生活中,人们可以假定某些命题总是正确的,除非有特别的关于例外的声明。比如,甲对乙说:“我要送你一只小猫”,第二天,甲真的拿来一只小猫,但却是一只死猫。这时乙当然会不高兴,“你说要给我一只小猫,为什么给我一只死猫?”于是,甲就申辩说,“我可没有说过要给你的猫是死的还是活的,猫可以是活的,也可以是死的。”问题出在哪里?问题在于:作为日常的语言交际,我们有一些心照不宣却必须遵守的规约(convention),这种约定是不用说出来的,但又是大家心知肚明、不可违反的。比如,说给人一只猫或鸟,肯定是活的;如果是死猫或死鸟的话,那么就必须事先声明:“那可是一只死猫唉,你爱要不要。”这是一种交际常规(communication regulation)。这种推理方式不光在日常生活中管用,就是在法庭上也是有效的。比如“人工智能”一词的创造

者、斯坦福大学的 McCarthy 教授举了这样一个例子：一个主人请了一个木匠给他做一个养鸟的棚。做完以后，主人拒绝付钱，理由是：你给我盖的鸟舍是有棚顶的，而我要养的是一只鸵鸟，根本用不着这个棚顶，鸵鸟又不会飞，所以拒绝付工资。结果争执不下，就告上了法庭。法庭判木匠胜诉，因为如果你养的是一只鸵鸟，不要棚顶，那么你就应当事先声明。的确，一般人听到鸟，总是会想到它会飞；也就是说，[会飞]是鸟的缺省性特征，是不用特别说明的。

同样，当我们说“房子大”的时候，[面积]这个特征是缺省的；当我们说“箱子大”的时候，[体积]这个特征是缺省的。当我要表达房子的体积大的意思时，房子的“体积”这个词就不能省略。如“这座房子的体积真大，可以用来堆放棉花”中，“体积”这个词一般不能省略，因为[体积]不是“房子”的一个强特征。也就是说，[体积]虽然是房子的一个特征，但不是房子的缺省特征，它是一个有标记的特征(marked feature)，一定要说出来。同样的道理，当我们要说“箱子的面积大”时，箱子的“面积”这个词是不能省略的。这种推理方式就是缺省推理：只有强特征的词汇表达才能作为缺省的词，而弱特征的词汇表达是不能作为缺省的词的。所以人在进行语义理解时就是在这样一种神经网络、这样一种认知结构上进行这样一种缺省的推理过程。揭示这样一种过程，实际上就是揭示语义理解的微观的心理机制和逻辑机制，并为语义理解提供了一种可计算的逻辑模型。我的这一工作，发表在《中国语文》1994 年第 4 期上，名为《一价名词的认知研究》。后来清华大学计算机系的一位博士后研究人员，根据这个思想做了一个语义理解模型。他告诉我做得很好，机器上可以运转和验证。他还写了一篇文章，发表在新加坡的一个杂志上。

这是一种很有意思的工作，我们还可以用它来研究其他一些相关的问题。比如：

(3) 这酒很淡。{味道淡 > 颜色淡}

(4) 这花很淡。{颜色淡 > 味道淡}

它们的意思不一样。说“这酒很淡”，只能是指这酒的味儿很淡；说“这花很淡”，只能是指这花的颜色很淡。如果我们认真地想一想，就

会觉得酒也有颜色;但是,当我们要表达{这酒颜色很淡}这种意思时,"颜色"这个词是不能省略的,一定要老老实实地说"这酒颜色很淡,但是喝起来很冲,味道很浓"。同样,如果要表达{这花的味儿淡}这种意思时,"味儿"这个词也是不能省略的,一定要老老实实地说"这花儿味儿很淡,虽然看上去很艳,颜色很浓"。为什么?我们可以用刚才说的那一套办法和概念,就是语义网络、扩散性激活、缺省推理等来解释。它还能解释一批其他相关的复杂的语言现象,涉及到人类处理语言形式和意义、看到的语言形式和脑子里想到的意义之间的关系的微妙复杂的机制。这些问题如果用传统的理论就不好解释。现在我们从认知心理学、从计算机科学的角度来研究这些问题,就不光能得到一个很好的解释,并且形式化的程度很高,刚才那些认识可以用一些产生式规则(production rule)来描述,从而可以算法化,并通过程序语言在机器上实现。

从这里我们看到,语言学研究一定要面向当代的科学技术,这样才能获得有用的概念和合适的研究方法;同时,这种研究结果才能真正为现代科技服务。这是一种语言研究和当代科技的双向互动,而不仅仅是单纯的语言研究。

2 语义演算和名词配价

下面讲第二个案例,关于语义演算的问题。

这个问题很早以前就是一些哲学家、数学家、逻辑学家非常关注的。他们希望通过一些像数学演算一样的方法,来证明这些句子是可以说的、那些句子是不能说的,这个句子是有意义的、那个句子是没有意义的。比如:

(1) 王冕死了。

要判断这个句子到底有没有意义,如何判断?逻辑学家说,有无意义的问题就是有无真值的问题。说一个陈述有真值,就是说这个陈述是可以判断真假的。那么,如何来演算呢?这就首先要引入一些范畴表达式,把句子中的词项及其语法关系进行形式化表示。我们用

e来代表“王冕”的所指（e是英语entity的缩写，表示实体），把“死”的意义表示为〈e，t〉，即从e到t的一个函数（t是英语truth的缩写，表示真值）。体态助词“了”暂时不考虑。这样的符号表示，对大家来说可能不好理解。现在，我们牺牲掉一点精确性，尽量把它说得直观一些：“死”这个词本身表示的意义是没有什么真假可言的，只有当引入一个实体词跟它组合时，才能判断由它组成的这个陈述是真是假，这就是函数表达式〈e,t〉的含义。比如，光说“死”，你不知道其真假；只有说“张三死了”、“李四死了”时，才能知道这个陈述是真是假，单独一个动词“死”就无所谓真假。拿“王冕死了”这个句子来说，假如“王冕”是现实生活或某个可能世界（possible world）中的某一个人；在现实生活或这个可能世界中，如果真的王冕死了，那么这句话就是真的；如果王冕没有死，那么这句话就是假的。不管是真是假，这句话都是有意义的。如果一个句子说出来后不能判断它的真假，那么这个句子就是没有意义的。这种逻辑就是逻辑学家蒙塔古（Richard Montague）的内涵逻辑（intensional logic）。在这种逻辑体系下，“王冕”属于范畴〈e〉、“死了”属于范畴〈e,t〉，于是，“王冕死了”可以通过逻辑演算来看它有没有意义（即真值）。例如：

王冕　死了。
〈e〉　〈e，t〉
———————
t

把两个范畴表达式中的e约分以后，最后剩下的结果是t，说明这个句子是有真值的。我们可以把范畴表达式〈e,t〉解释为：它是一种小型的计算装置，只有向它输入一个实体e，它才能输出一个真值t。这里正好有实体词“王冕”出现，所以约分以后得t，就是说这个句子是可以判断是真或假的。这就是蒙塔古语法的演算方式，非常严格、非常精致、非常漂亮。问题是，这套办法拿到汉语中来后，对有些句子是不大好处理的。比如：

(2) 王冕七岁上死了父亲。

死的不是王冕，而是他父亲，能不能再用上面那套办法来推导？不好推导。如果这样计算：

王冕 父亲 死了
〈e〉〈e, t〉〈e〉
t
?

t 和 e 对不起来，剩下的不好处理。如果先算后面的两个范畴也一样：

王冕 死了 父亲
〈e〉〈e, t〉〈e〉
t
?

也没有真值。但上面这句话却肯定是合语法、可接受，并且是有意义的。怎么办？原来“父亲”这类词不能简单地跟“王冕”等同起来作为一个实体，这是不对的。因为“父亲”是一个关系词，单说“父亲”是没有所指的，一定要说成“小王的父亲”（是指老王）、“小张的父亲”（是指老张）之类的形式，才能获得所指意义。所以“父亲”不能这样简单地用范畴表达式〈e〉来刻画。我们给这类名词取一个名称，叫做“一价名词”。意思是，它需要一个配价成分跟它组合，才能有真正的所指。在语言中，大多数名词是零价的，不需要配价成分。少数名词是一价的，需要一个配价成分，比如“尾巴、抽屉”等，都需要另外一个名词跟它配合，说成“狐狸的尾巴、书桌的抽屉”等，才能有语义所指。还有少数是二价的，需要两个配价成分，比如“意见、感情”等，都需要两个名词跟它配合，说成“学生对老师的意见、大李对刘芳的感情”等，才能有语义所指。在上例中，“父亲”是个一价名词，它需要另外一个实体词的出现，才能有所指。所以这里应当把它表示成一个函数式〈e, e〉，大意为：它是一个从实体到实体的函数。直观的理解可以是：这种词表面上看起来好像表达一个实体，其实这是一种抽象实体，它一定要另外一个实体出现时，才能有所指，才能真正表示可能世界或现实生活中的某一个人、某一个事物，等等。比如，光说“哥哥”、“父亲”是没有所指的，不知道是谁，一定要说“谁的哥哥”、“谁的父亲”才行。所以它是一个〈e, e〉式的函数。这样一来，上例的语义演算就好处理了。例如：

王冕 死了 父亲

$\langle \cancel{e} \rangle \langle \bcancel{e}, t \rangle \langle \bcancel{e}, e \rangle$

$\underline{\qquad\qquad\qquad}$

$\langle \cancel{e}, t \rangle$

$\underline{\qquad\qquad\qquad}$

t

上述约分表达式,可以这样来直观地理解:“死了”跟“父亲”先组合,各消掉一个实体以后,整个“死了父亲”在功能上相当于一个不及物动词,仍需要另外一个实体词填入,才能构成一个可以判断其真假的陈述。当然,这是一种大概的设想。从逻辑技术上讲,其中的贴合运算的步骤和规则等细节,还尚待作出进一步的研究。这样,“王冕死了父亲”是有真值可言的。不管它是真是假,这个句子是有意义的。这就是一种语义的演算。但这是简单的,还有复杂的,比如在上例中加上一个副词“刚刚”,说成“王冕刚刚死了父亲”,你就得引入更为复杂的演算规则。如果所有的句子都能够用这种办法来演算的话,那机器完全可以根据这种办法来自动地推导:一查词典,“父亲”是这样一种范畴,然后自动给它标上相应的范畴表达式,再对有关词语的范畴表达式进行自动演算,就可以自动判断句子是有意义的还是无意义的。能按照规则推出 t 的句子,就是有意义的;不能按照规则推出 t 的句子,就是没有意义的。

这是用现代逻辑的方法来研究语言,这种方法对机器处理语言比较有用。我的同学白硕在中科院计算所工作,他基本上就是用这种办法建造了一个句子意义的演算系统,希望用于信息检索、快速查找等语言信息处理工程。在他的演算系统中,他采纳了我提出的关于名词配价的有关思想。这对语言学者来说,是比较兴奋的一件事。

以上两个案例听起来是比较费脑子的,下面我们来两个轻松一点儿的。

3　范畴辨认和词类划分

第三个案例,范畴的辨认。

生活中需要辨别“类”跟实体的关系。哪些实体可以归入一个类,哪些不能归入一个类,这就叫范畴的辨认。心理学家很关心这么一个问题,做了很多实验。我们用它来研究汉语的词类。在座的各

位可能都有这样一种经历：买一本《现代汉语词典》回来，翻开来一看，发现词典上没有标名词、动词、形容词等词类，顿时觉得很遗憾。相应地，你买一本再小的英语词典，上面都标明 verb 或者 noun、adjective。为什么汉语词典不标词类？是不是汉语的词没有词类？不管你回答"是"还是"不是"，都需要证明。我们不妨先假定汉语的词没有词类。我的同学白硕曾经这样论证过：如果说汉语没有词类，那么从逻辑或数学的角度看，只能有这样两种情况：

第一种可能性是：所有的词都是一个类，没有分别。假设有 A、B、C 任意三个汉语的词，如果没有词类分别，那么在下面这个句子框架中，

主	谓	宾
a	b	c

A 就既可以出现在 a 的位置上，也可以出现在 b 的位置上，也可以出现在 c 的位置上；B、C 也都可以出现在 a、b、c 三个位置的任意一个位置上。就是说，任何一个词都可以出现在句子的任何一个位置上。这种情况是不符合汉语的实际的，我们明显地感觉到"的"、"着"、"了"这样一些词总是出现在其他成分的后面；而"把"、"被"、"也"这样一些词总要加在某些成分的前面。由此可见，第一种假设是不成立的。

第二种可能性是：所有的词都两两相异，各不相同，谁跟谁都不一样。因为每一个词都自成一类，所以也无词类可言。但这也不符合语言事实。例如：

V___ O

在这个空格中可以加"了"、"着"、"过"，至少这三个词是一个类。这样看来，汉语的词类也不可能是这种情况。

我们排除了上面两种假设，那么汉语的词类就应该是有的了。下面我们来讨论汉语有词类的情况。

我们首先会想到名词、动词、形容词，但问题是，我们有没有办法来准确地识别哪些词是名词，哪些词是动词，哪些词是形容词？这种办法还是有的。自从结构主义在上个世纪二三十年代兴起以来，形

成了一种分布分析(distribution analysis)方法。有些词如果所处的句法位置是一样的，那么它们是一个类。比如有几个词既能作主语，还能作宾语、定语，这些词就是一个类，叫做名词。有的词能够做谓语，一般不能直接单独作主语、宾语、定语，这些词可以叫动词。这种分析方法就叫做分布分析。严格地讲，用分布分析划分词类时，应该考虑一个词在语法上的全部分布(total distributions)；只有全部分布都相同的词才能归为一类。但是，一个词的所有的分布太复杂、不容易穷尽，并且真正全部分布都一样的词是不多的。那么简单化一些，怎么办呢？我的老师朱德熙先生的做法是：找一些关键的特征，即一些区别特征，利用这些特征之间的合取或析取关系，作为划分词类的标准。根据他的说法，名词是适合于下列分布框架的一类词：

N：SL ___ & *F ___

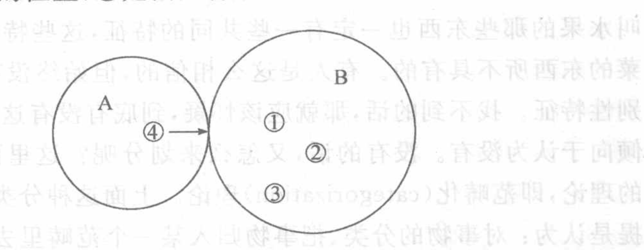
即它通常受数量词(记作 SL)的修饰，同时不能受副词(记作 F)修饰。比如，不能说“很桌子”、“太桌子”。这就是找出了具有区别性的分布特征。这个办法非常好，动词、形容词也可以用这种区别性的分布特征来划分。这种方法在原则上可以对汉语的词类问题进行很好的处理，但还是有很多问题不能很好地解决，比如说，有些词，在感觉上完全是一个名词，可是它适应不了这个框架。比如“体育”、“皮肤”。“一项体育”说不说？不说。“一个体育”也不说。“皮肤”呢？说“一块皮肤”的可能性很小很小，不是完全没有可能，外科大夫有可能在手术室里说“一块皮肤”。一般情况下，“皮肤”也不受数量结构修饰。按照朱先生的这个框架，就把它们从名词中排除出去了。但我们却能明显地感到“体育”和“皮肤”肯定是名词。如果“桌子、椅子”是名词的话，那么“体育、皮肤”也应该是名词。看来，这一套办法要受到怀疑：分布分析，或者找区别特征的办法可靠性到底有多大？能解决多少问题？更为极端的想法就会怀疑：汉语的词到底能不能分出词类来？汉语到底有没有词类这种语法范畴？

我们从认知心理学的范畴辨认方面来分析它。下面我们先不考虑词类，而是考虑日常生活中的问题。比如说蔬菜、水果的分类。现在我们要想一个办法，找出一个定义来，什么是蔬菜，什么是水果？

并且根据这个定义，把应该属于水果的，归入水果，把应该属于蔬菜的，归入蔬菜。能不能？比如说，水果是木本植物的果实，可以直接吃的；蔬菜是草本植物，炒了以后才能吃的。这样下定义行不行？好像不行。因为甘蔗也是水果，但它不是木本植物的果实；草莓也不是木本植物，但它却是水果。但是我们会不会怀疑蔬菜和水果这种区分是无效的呢？或者声称：根本就没有蔬菜、水果这种分类。显然不会。那么，为什么我们刚才会怀疑名词、动词、形容词这些范畴的存在，而现在却不怀疑水果与蔬菜是两个客观存在的范畴呢？我们是经过了哪些认知的操作来做到认识哪些东西是蔬菜，哪些东西是水果的呢？这是搞心理学的人很关心的问题。有的人想，只要锲而不舍地追下去，总能找到一些特征，把它们分开。是蔬菜的东西一定有一些共同的特征，这些特征是那些叫水果的东西所没有的；同样，叫水果的那些东西也一定有一些共同的特征，这些特征是那些叫蔬菜的东西所不具有的。有人是这么相信的，但始终没有找到这种区别性特征。找不到的话，那就应该怀疑，到底有没有这种特征？我们倾向于认为没有。没有的话，又怎么来划分呢？这里面涉及到分类的理论，即范畴化(categorization)理论。上面这种分类的概念，其前提是认为：对事物的分类、把事物归入某一个范畴里去，根据的是事物的特征；这个特征是一种充分必要条件，是这个范畴的成员一定要具有这个特征，并且不属于这个范畴的东西一定不具有这个特征。这是一种“是”或“否”的简单的特征分类。根据特征分类得出来的范畴是特征范畴，数学上的很多范畴都是特征范畴。比如说奇数和偶数，那是很清楚的，如果偶数，一定能被2整除；如果奇数，肯定不能被2整除，很清楚。又如素数，必定只能被1和它本身整除，是这个范畴的，一定有这个特征；不是这个范畴的，一定不具有这个特征。那么，汉语的词类是不是属于特征范畴呢？不是。用这种特征分类来分词类，就行不通。我们下面考察蔬菜、水果这种类是怎么分出来的。

认知心理学认为，蔬菜和水果的分类是根据一种原型分出来的，这种范畴叫做典型范畴或原型范畴(prototypical category)。它的意思是：在构成这个范畴的众多成员中有一批成员是核心的成员

(即原型),这些成员都具有某个区别于其他范畴的共同特征。其他的成员是边缘性的成员,它们是通过与典型成员相比较,根据它与典型成员的相似性来归入的。所以一个范畴里面可以分成典型成员和非典型成员。原来,典型的水果确实是木本植物的果实,并且确实是可以拿来就吃的;其他的一些成员跟它类比,或者是拿来就可以吃的,或者是木本植物的果实,或者有其他一些共同因素,通过比较类比的方法归进来的。典型的蔬菜是草本植物,炒了以后才可以吃,如青菜、菠菜,这是典型成员;其他有些成员根据类比,如西红柿,就是由于是草本植物的果实而归入蔬菜的。如果跟水果相比,它也具有拿来就能吃的特点,因而归入水果也可以,所以西红柿处在两个范畴的交界处。如下图所示:A是蔬菜,B是水果。①、②是苹果、梨等,甘蔗则处在③的位置,④是西红柿。



用这个思想来看汉语的词类,我们就豁然开朗了。为什么说汉语有词类,却又分不清楚;说它没有,又不能令人信服?想想“桌子”肯定是名词,“吃”肯定是动词,“好”、“饱”肯定是形容词,怎么汉语会没有词类呢?肯定有。要说有,可是又分不开来。问题在于原来是用“特征范畴”这样一种眼光来看问题的,戴错了眼镜,所以看东西看不清楚。现在有了典型范畴这种思想,我们就可以用它来看这个问题:原先朱先生给出的这两个特征,我们可以把它看成是名词的典型成员的分布特征,其他成员跟这些典型成员进行类比,相似性比较多的话,算作名词,比如“皮肤”跟“体育”,都不能作谓语,只能作主语、宾语、定语;尽管不受数量结构修饰,也可以算它是名词。用这种新的眼光来看,那么我想汉语的词类是可以解决的。我这套思路提出来以后,很多人都不能接受,觉得总是把握不准。后来我有个同学,他是在中科院的计算所工作的。他看了以后很高兴,说在计算机

上可以模拟。在人工智能上有一个专门的学科，叫做 Machine Learning，即机器学习；让计算机跟人一样学习，让计算机更加聪明。机器学习有很多方法，其中有一种叫做“通过类比的学习”(Learning by Analogy)。用这套方法，先去告诉机器典型的名词有哪些特征，机器通过类比学习，可以把名词、动词、形容词大致地区分开。

讲到基于类比的学习，我们介绍国外人工智能界做的一些工作：让机器看剧本，自动地找出哪些是喜剧，哪些是悲剧。如何实行？先输入《罗密欧与朱丽叶》这么一个剧本，对它进行分析：罗与朱相爱，感情非常真挚而且异常强烈，然后发生了一连串轰轰烈烈的家族争斗，结局是有情人难成眷属，并且都死于非命，天崩地裂之后白茫茫一片大地真干净。研究者根据各种关系制成图形，表示这种关系并输入机器。接下来再输入一个《奥赛罗》的故事，机器把它与《罗密欧与朱丽叶》相类比，把它分析成悲剧。科学家还让机器去自动地发现欧姆定律。先教机器去分析一个水管，把流过一段水管的水流的速度、压力、摩擦力等参数和它们之间的关系都告诉它，机器知道了这些以后，就会自动地类推：电流、电压、电阻是否也有这种关系？如果有的话，就会得出这样一个定律。这样，机器就像人一样地发现了欧姆定律。甚至有更极端的做法，把这种方法推广到所有具有线性的约束关系的系统，可以用这种机器学习的办法去发现更多的科学定律。我们的研究就是在这样一种背景下进行的。它的意义和价值不光是语言学的，它还与人工智能有关。

4　认知策略和语序排列

最后讲语序排列的问题。我们也从认知的角度去研究它。先看下面的例子：

小红球　　大木盆　　小黑铁塔
*红小球　*木大盆　*黑小铁塔

名词前面有一些定语，它们的排列是有规律的，这里我们不考虑加“的”的情况。“红球”可以说，“小球”也可以说，但是加在一起就只能

说“小红球”，不能说“红小球”。“大木盆”、“小黑铁塔”也是同样的情况。这就是定语的一种排列顺序。我们有没有可能把这种规则提取出来。这个工作其实十几年前就有人做了，直到现在还不容易说清楚，很复杂。我们假定有一种观察方式，从语义上来考察，然后给出一条规则（>表示“先于”）：

R_1：尺寸 > 颜色 > 质料

用这条规则来解释上面的例子，是可以控制的；但是加一些例子进来，就覆盖不住了。比如，“农民棉花专家”，不能说“棉花农民专家”，这里面又有它自己的规则，用 R_1 控制不了。再如“当代青年化学家”，“当代”和“青年”都与时间相关，但两者的次序是固定的，不能说“青年当代化学家。”可见 R_1 很直观，但是预测能力很差，不能解释更多的问题，所以这种分析是浅层次的分析。到目前为止，这个问题之所以没有能很好地解决，关键在于大家的眼光始终盯在意义上：“小”、“大”表示尺寸，“红”、“黑”表示颜色，“木”、“铁”表示质料，跳不出这个圈子，所以这个问题解决不了。现在我们要换一个角度来研究它，不单纯从语义类，而是从语义聚合的角度来观察。这样，我们可以发现，“小”是跟“大”对立的，最多再跟“中”对立，形成“大”“中”“小”三项对立。“红”的对立项就有“蓝”、“白”、“黑”等至少七项。而“木”、“铁”等表示质料的对立项就更多了。到了这时候，定语排列的语序规则就很清楚了。可以表示如下：

R_2：对立项少的定语 > 对立项多的定语

R_2 比 R_1 抽象，同时，预测能力增强了。并且 R_2 也可以对 R_1 进行解释，为什么表示尺寸的定语要放在表示颜色的定语前边？因为尺寸的对立项少，只有大、中、小三种，而颜色的对立项多，所以放在后面。R_2 比 R_1 要好得多，或者说更有洞察力（insight）。

但是为什么要这样排列？一条规则出来以后，如果不能对它作出解释，那么它就是不完美的，缺少理论的魅力。所以我们要作出进一步的解释。

我们从信息量的角度来解释，依据下面的信息理论：一个信号所传递的信息量的大小，并不依赖于这个信号本身，而是依赖于跟这

个信号能够构成替换关系的信号的数目。如果能够替换的数目很多，那么这个信号的信息量就很大；如果能替换的很少，那么这个信号的信息量就很小。我们容易理解这个。比如英语考试的时候，不会考这种填空题：

exi $\underline{x}$

因为，这里 x 处只能填 t。请问，这个 t 有多少信息量？显然，它几乎不传达信息，因为这儿只能填 t，答案是唯一的。但是换一个词：

rea $\underline{x}$

情况就大为不同了。因为，这里 x 处可以填入 l、m、p、r、d 等，从而造成 real、ream、reap、rear、read 等单词。于是，这里面每个 x 的值传递的信息量就很大。再举一个例子，假如我们只有一个红球放在口袋里，不用猜就知道这是个红球，没有什么信息量。如果一个红的一个白的在口袋里转几圈，拿出一个来猜，那么信息量就增大了，猜中的可能性就减少了。所以如果从信息量的角度看，我们可以把 R_2 改写成 R_3：

R_3：信息量小的定语 ＞ 信息量大的定语

事实上，R_3 是对 R_2 的解释。如果把语言看成是一个自足的系统，即完全独立的系统，那么 R_3 就到底了，这是语言内部的规则，跟外部不发生交流，纯粹是一种独立的语言信息的组织和安排方式。但语言只是人类认知系统中的一个子系统，并且语言是通过认知过程来产生和理解的，这中间有一个认知加工的问题。具体地说，人们说话时，经过大脑认知加工产生语言。听到话语时，用大脑加工得到它的语义。所以，这条规则是应当从认知的角度作出解释：为什么信息量小的定语在前，而信息量大的定语在后？

这里面涉及到人类信息加工时的基本策略：先处理简单信息，后处理复杂信息。举个例子来说，考试的时候，发下试卷来，一看前面的八个题目是容易的，后面两个问题是难的，每题 20 分，前面八个大题才 60 分，你会采取什么策略？先做简单的，后做后面 40 分的难题。如

果纯粹从理性的角度看，这是不合理的，刚考试的时候，头脑清醒，应该先做难题，把40分拿下来，然后随便做几道，60分就到手了。但是大家偏偏不这样做，我们往往先用很多时间去做大量的填空题呀，名词解释呀，最后才做难题。为什么？我们脑子里根深蒂固的策略是先处理容易的信息，后加工复杂的信息。已经形成了一种思维定势。

一个句子可以从语用上分成话题部分和说明部分，话题传达的一般是旧信息，说明传达的一般是新信息。比如，我们在北京坐地铁的时候，听到地铁上用英文报站名：

The next station is *Xizhimen*. *Xizhimen* is the next station.

为什么不倒过来说成：

Xizhimen is the next station. The next station is *Xizhimen*.

为什么？因为在坐车的语境中，The next station 是个变量性的成分，没有明确、绝对的所指，语义信息量很小，容易加工，而 *Xizhimen* 是个绝对性的地点词语，传达一种新信息。已经有了前一句中 *Xizhimen* 这个确定的信息以后，再说 *Xizhimen* is the next station. 就很自然。讲故事也是这样，我们通常把新信息放在后面，慢慢地引导出来，然后再放在另一个故事的开头。比如，上一回评书的结语是

……四位大汉正在酒楼上边喝边聊，指手画脚地议论着南拳北腿、东邪西毒，谁是天下第一大侠。咚！咚！咚！忽听得楼梯上一阵脚步声，众好汉吓得一个个脸色如土。要问来者是谁？且听下回分解。

害得听众夜里都睡不好觉，脑子里老在琢磨：来的是哪位武林高人，竟然让四位英雄都大惊失色。第二天，一听下一回的开头却是：

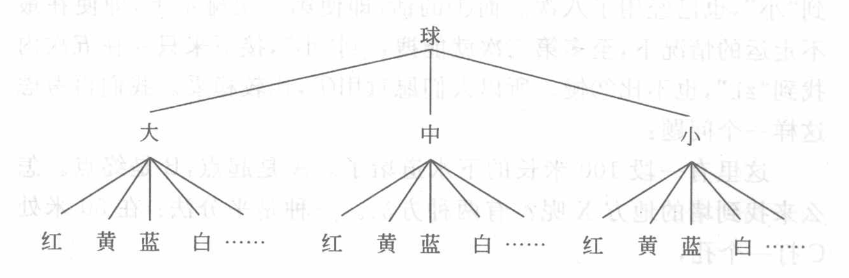
前文书，我们说到……。来的不是别人，原来是上菜的伙计……。

这也是为了适应听众信息加工的心理需要：让他们先温习已经知道的故事情节，先加工这些简单的已知信息；然后，再交代他们最关心的新的故事情节，解开上次留下来的悬念，加工这些相对来说是复杂

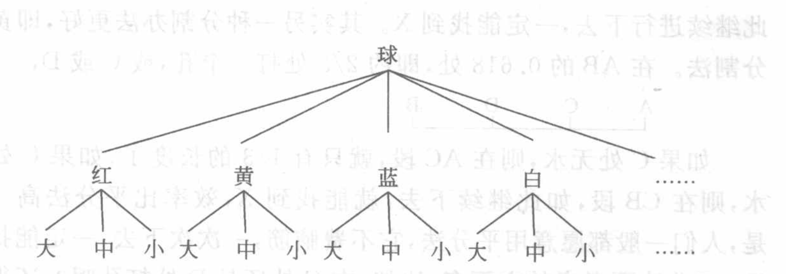
的新信息。

比如，上文说的定语顺序问题，还可以用心理学和计算机科学上的搜索(search)理论，来作出如下这种解释：

①"小红球"的搜索结构：



②"红小球"的搜索结构：



我们把修饰定语的中心词看成是一个问题空间，找某一种颜色、某一种尺寸的球，这是一个目标。为了搜索到"小红球"这个目标，我们把它分成几个组，组下又分层次。不同的是先分得多些，再分得少些；还是先分得少些，再分得多些？这两种表达方式构成两种不同的搜索路径，同时也构成两种不同的问题求解的空间结构。这两种解决问题的空间结构是不一样的，在心理学和人工智能上，通常把①叫做深度优先策略(depth first strategy)，把②叫做广度优先策略(breadth first strategy)。从概率上讲，两者是等价的；从搜索的效果上讲，②的效果更好些。因为如果它碰巧第一次搜索到"红"，接下来最多搜索三次就能找到目标。第一种方式即使第一次搜索到"小"，接下来的可能性还是不大。但是人们还是喜欢用①这种形式，

先找“小”，再找“红”。也就是说，人们在信息加工时，乐于用深度优先的策略，而不用广度优先的策略，为什么？因为尽管②的效率是高的，但是人们想万一第一次碰不上，搜索“红”可能会需要很多次数，比如有七种颜色的话，就有可能要搜索七次。接下来即使一下子碰到“小”，也已经用了八次。而①的话，即使第一次碰不上，即使在最不走运的情况下，至多第三次就能搜索到“小”，接下来只要在五次内找到“红”，也不比②慢。所以人们愿意用①，比较稳妥。我们再考虑这样一个问题：

这里有一段 100 米长的下水道堵了。A 是起点，B 是终点。怎么来找到堵的地方 X 呢？有两种方法。一种是平分法：在 50 米处 C 打一个孔，

A C B

如果不流水，那么 X 在 AC 段；如果流水，那么 X 在 CB 段。如此继续进行下去，一定能找到 X。其实另一种分割办法更好，即黄金分割法。在 AB 的 0.618 处，即约 2/3 处打一个孔，或 C 或 D，

A C D B

如果 C 处无水，则在 AC 段，就只有 1/3 的长度了；如果 C 处有水，则在 CB 段，如此继续下去，就能找到 X，效率比平分法高。但是，人们一般都愿意用平分法，它不费脑筋，一次次下去，一定能找到 X。三分法要考虑的方面多，比如，在 C 处还是 D 处打孔呢？还得琢磨一番，去计算也麻烦，大脑不愿意去做这种选择工作，这是人类信息加工的一个基本策略。因此，全知全能的哲学家要抱怨说：人是一种具有有限理性的动物。

我们可以把 R_3 推广，把“信息量小的定语”改成“信息量小的成分”，就得到下面的规则：

R_4：信息量小的成分 ＞ 信息量大的成分

并且还可以继续追究：为什么是这种顺序呢？因为信息量小的成分容易加工。于是，得到下面这条更抽象的规则：

R_5：容易加工的成分 ＞ 不易加工的成分

这一套规则基本上是语言中的普遍规则。因为人类的认知结构相似，人类的认知过程、认知策略也差不多。

当然，语言里面语序的排列除了遵循这套规则以外，还有其他一些规则。有些现象是不能用这套规则来解释的，因为不同的规则之间要相互打架、竞争。比如有一个句子不符合信息量规则：

白色长统袜子

长统、中统、低统，最多三个对立，白色却有很多对立。按照信息量原则，应当说成“长统白色袜子”，可是这不行，不合语法，非得说成“白色长统袜子”不可。为什么？原来，颜色词修饰名词是受限制的，“白色袜子”不能说，要说“白色的袜子”。“白色（的）长统袜子”中的“的”实际省略了。为什么能省略？名词的定语比较多，层层修饰的时候，“的”可以省略；单独一个定语，“的”不能省略。这样，语言内部规则要求“白色”跑到“长统”的前面去，且可以省略“的”。

我举了这么一些例子，来说明我们可以从认知科学，尤其是信息加工心理学这个角度来研究语言；同时，还从计算机科学和技术的角度来研究语言。这样的语言研究的成果可以反过来对计算机科学、心理学和其他科学都有实用的价值，甚至对于对外汉语教学也是比较管用的。以上举的例子都是比较简单的，更进一步的问题由于时间关系就不再介绍了。有些可以参看《国外语言学》1996 年第 2 期上我写的一篇介绍认知语言学的文章。

好，今天我就讲这些问题，谢谢大家！

1996 年 10 月在北京大学中文系“孑民学术论坛”上的演讲
徐　刚　记录整理
（收入费振刚、温儒敏主编《北大中文研究》，
北京大学出版社，1998 年）
2004 年 9 月改写

二、论元结构和描述框架

论元角色的层级关系和语义特征

本文主要讨论汉语动词的各种论元角色的层级关系，详细地刻画各种论元角色的动态性的语义特征。首先，介绍国外语言学界对于论元、论旨角色、论元位置、论元结构、论旨阶层等概念的定义和认识。然后，讨论汉语动词的论元结构研究的五个方面的内容（论元数目、论旨角色、句法特征、语义特征、配位方式）及其基本的研究原则和处理技术。接着，讨论不同的论元角色之间的关系，建立一个汉语动词的论元角色的层级体系；着重描写各别论元角色在述谓结构中所表现出来的动态性的语义特征，也兼及它们各自的句法特征。最后，通过实例讨论不同的论元角色在句法上的共现关系和语义上的转化关系。

1 引 言

对现代汉语动词的论元结构进行研究，一方面可以发现施事、受事等语义成分跟主语、宾语等句法成分之间的投射关系，加深我们对汉语的结构面貌的全面认识；另一方面可以为计算机处理汉语提供比较充分的语义知识方面的资源，满足机器翻译、信息抽取、快速检索等涉及语义信息的处理技术的需求。当然，汉语动词的论元结构方面的知识，可以帮助第二语言学习者更好地理解汉语的句子结构和语义解释之间的映射关系，对于对外汉语教学也有直接的应用价值。有鉴于此，我们开展了现代汉语动词的论元结构的研究项目。在此，我们首先把在研究过程中碰到的一些原则性的问题摆出来，说明自己的观点和处理办法，然后着重讨论汉语动词的各种论元角色的层级关系，详细地刻画各种论元角色的动态性的语义特征、也兼及其句法特点，通过实例讨论不同的论元角色在句法上的共现关系及其论旨角色转换的语义机制。以期起到抛砖引玉的作用，希望得到广大同行的指正。

2　关于论元结构的几个基本概念及其含义

要研究汉语动词的论元结构，首先必须引进几个理论概念，并严格厘定其含义和使用范围。为了便于跟国外的相关研究进行比较，我们首先根据顾阳(1994)，介绍下列术语及其定义：

(1) 论元(argument)：指带有论旨角色的名词短语。

(2) 论旨角色(thematic role)：由谓词根据与其相关的名词短语之间语义关系而指派(assign)给这些名词短语的语义角色。谓词有其固有的论旨角色，这些角色表示谓词所涉及的主体、客体或动作、行为、状态、所处的场所、动作的起点、方向、终点、原因及引起的结果、凭借的工具，等等。目前公认的论旨角色有施事者(agent)，感受者(experiencer)，受惠者(benefactive)，客体(theme)，使役者(cause/causer)，等等，并通常将受影响的客体称作受事者(patient)。

论旨角色这一概念的产生及运用反映出语言学家试图透过表层的语法关系，如主语、宾语同述语的结构关系，更深入地了解述语与论元成分之间的语义关系，以及这种语义关系对语法的影响。

(3) 论元位置：论元在句中所占的位置。

(4) 论元结构(argument structure)：一个词项的论元结构就是该词项所能拥有的一组已经标有论旨角色名称的论元。这是把论元结构看作是论旨角色关系的同义词，论元结构中所含的内容无非就是一系列的论旨角色。

(5) 论旨阶层(thematic hierarchy)：指论旨角色在词汇概念结构中的排列形式，由于大家相信论旨角色是按照阶层的形式排列的，因而称为论旨阶层。例如：

施事 > 处所/终点/起点 > 客体

论旨角色在论旨阶层中的位置跟其在句子中的位置(即论元位置)直接相关，比如：施事通常占据主语的地位，处所等通常占据状语的地位，受事通常占据宾语的地位。

从上面的术语解释和说明，我们大致可以看出国外论元结构研究的主要内容和目的。

3 动词论元结构研究的主要内容和相关原则

汉语动词的论元结构的研究应该吸收国内外配价语法、格语法、生成语法、论元结构理论研究的有关成果，特别是最近二十多年来汉语动词的配价研究的成果，根据计算机处理汉语等实际应用的需要来确定汉语动词的论元结构的研究内容。其主要内容应该包括：①

(1) 论元属性：确定每一个动词能支配多少个必用论元、多少个可用论元；

(2) 论旨属性：标定这些论元在语义上的功能，即论旨角色；

(3) 语法特征：描写这些论元在句法上的功能和所受到的句法约束；

(4) 语义特征：描写这些论元的动态的语义特征和静态的语义特征；

(5) 配位方式：描写动词及其论元的句法配置方式。

"论元属性"(argument property)指的是动词所能关联的论元的数目，这方面的内容可以参考配价语法的研究成果；"论旨属性"(thematic property)指的是各个论元的论旨角色，这方面的内容可以参考格语法的研究成果；"语法特征"(grammatical feature)包括句法功能(syntactic function)和范畴特征(categorical feature)两个方面，前者指各个论元在句子中各自可以充当什么样的句法成分(比如：主语、宾语、状语)，后者指各个论元通常由什么样的词类范畴来实现(比如：施事、受事通常由名词性成分来实现，致事通常由名词或动词性成分来实现，场所、源点、终点通常由处所性成分来实现)，

---

① 参考汤廷池、张淑敏(1996)，第261页。

这方面的内容可以参考论元结构的研究成果;“语义特征”(semantic feature)包括动态的语义特征和静态的语义特征两个方面,前者指的是各个论元在述谓结构中表现出来的施动性、受动性等语义特点,后者指实现不同的论旨角色的词语在语义上受到的约束(比如:施事、与事通常由指人名词来实现,受事则既可以由指人名词来实现,也可以由指物名词来实现),对于这方面的内容,可以参考词汇语义学和论元结构理论等的研究成果。“配位方式”(argument selection)指的是依存于同一个动词的各个论元在句子中的共现和选择限制,即怎样构成一个或几个相关的句式,这方面的内容可以参考配价语法和论元结构理论的研究成果。必须注意的是,所谓动词的论元结构实际上指的是动词的某个义项或义位(sememe)的论元结构;也就是说,同一个动词的不同义位可能具有各不相同的论元结构。

在研究这五项内容之前,必须首先明确下列问题:

(1) 怎样确定每个具体的动词所能支配的论元的数目,怎样区分必用论元和可用论元,在什么样的框架中确定动词所能支配的论元的数目,用介词引导的从属成分算不算数,怎样分清动词的论元结构和动词性结构的论元结构的区别?

(2) 怎样标定这些论元的语义角色,怎样处理论旨角色的模糊性问题,到底设立多少个论旨角色,怎样区分不同的论旨角色,要不要引进论旨角色的层级系统,怎样处理某些动词在特定句式中增加进来的论元?

(3) 怎样描写论元和谓词的句法配置方式,分不分基础句式和派生句式,标志性的虚词要不要包括进来?

(4) 怎样描写不同论元的语义特征,分不分动态特征和静态特征,不同论元的句法特征怎样抽取,怎样跟词义的层级体系和有关的句式描写挂起钩来?

最后,但并非最不重要的问题是研究结果的表述问题,这至少包括下列两个问题:

(1) 采用什么样的描写体例和表示方式?

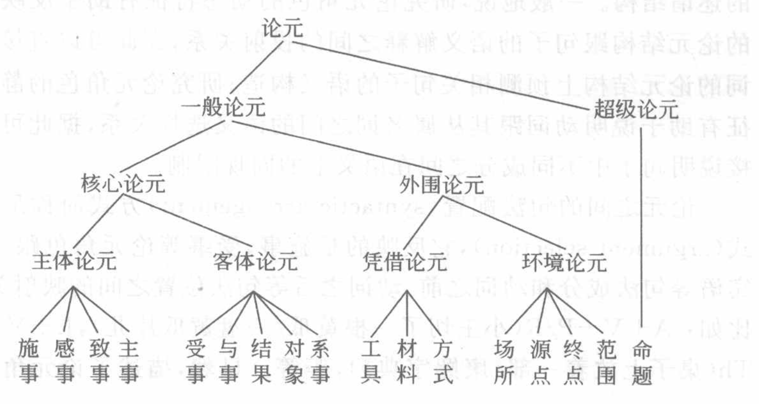
(2) 怎样处理说明文字、缩写符号、数字代码、通用规则、个

例说明、具体例证之间的排列关系？

关于动词的论元数目，我们的体会是：为了比较全面地反映动词对其从属成分的支配能力，为了最大程度上从动词的论元结构去把握有关句式的结构和意义，应该既包括必有论元(obligatory argument)、也包括可有论元(optional argument)。但是，限于那些从动词的词汇意义上可以推导出来，并且是在一定的句法结构中实现了的从属成分，形成一套相对可操作的、由动词所激活的意义场景跟“动词＋论元”式的句法配列(syntactic arrangement)相互验证的核查程序。这样，那些几乎可以跟所有的动词共现(co-occurence)的时间、处所等成分就应该排除在动词的论元结构之外。

关于论元的论旨角色(简称论元角色)，我们承认它源自人们对由动词所激活的语义场景(semantic scene)的认识，尤其是同一场景中相关的参与者(participants)之间的相对关系的认识。因此，论元角色难免有一定的模糊性、不容易用一种形式化的办法来严格地定义。但是，我们可以采用原型理论，首先给出不同的论元角色的典型的句法、语义特征，然后通过类比归类的办法来确定特征不明显的语义成分的论元角色。

必须指出的是，动词的各种论元角色可以根据其句法、语义特点而聚合成不同层级的类，从而形成一个论元角色的层级体系(hierarchy)。根据我们的经验，现代汉语动词的论元角色可以组织进如下这个层级体系之中：



其中,超级论元(super argument)指由谓词性成分充当的论元,这种论元本身就是某种论元结构的实现形式。核心论元(kernel argument)指动词的必有论元,它们对构成基本的述谓结构(基础命题)来说是不可缺少的;其中,主体论元以作主语为其主要的句法实现形式,客体论元以作宾语为其主要的句法实现形式。外围论元(circumstantial argument)指动词的可有论元,它们起到扩充基本的述谓结构、形成复杂命题的作用;它们以作状语为其主要的句法实现形式,其中凭借论元跟环境论元的区分主要出于语义上的考虑。

关于论元角色的语义特征,应该分为动态特征和静态特征两个方面。论元角色的动态特征,指不同类型的论元角色在述谓结构中所具有的语义特征,即在由动词所表示的事件结构中所表现出来的特征。比如,施事具有施动性、受事具有受动性、结果具有渐成性、与事具有自主性、工具具有移动性、材料具有变化性、处所具有不变性,等等。显然,这种特征不依赖于充当某种论元角色的名词性成分本身的词汇语义特征。论元角色的静态特征,指充当某种论元角色的名词性成分本身的语义特征;换句话说,具有某种语义特征的名词性成分比较适合于作某种论元角色。比如,作施事的名词性成分一般具有〔+有生(animate)〕和〔+人类(human)〕的语义特征、作工具的名词性成分具有〔+器具(tool)〕的语义特征、作材料的名词性成分具有〔+材料(material)〕的语义特征。显然,这种特征取决于充当某种论元角色的名词性成分本身的词汇语义特征,但是不依赖于语句的述谓结构。一般地说,研究论元角色的动态特征有助于反映动词的论元结构跟句子的语义解释之间的投射关系,据此可以直接从动词的论元结构上预测相关句子的语义构造;研究论元角色的静态特征有助于说明动词跟其从属名词之间的语义选择关系,据此可以直接说明句子中不同成分之间在语义上的同现限制。

论元之间的句法配置(syntactic arrangement)方式简称配位方式(argument selection),它反映的是施事、受事等论元角色跟主语、宾语等句法成分和动词之前、动词之后等句法位置之间的映射关系。比如:A+V+P/R(小王切了一根黄瓜/一盘黄瓜片儿),L+V着+Th(桌子上放着一部《康熙字典》),等等。显然,描述了论元角色的

句法配置，就等于是建立了一种标注了语义关系的句法格式；这种句法格式当然是一种最接近语义表达式的句法表达式，便于建立句法结构形式跟语义结构关系之间的映射和连接。配位方式的研究，应该做到以下四点：

(1) 要充分反映不同的论元角色之间的同现限制及其在表层结构中的语序位置；

(2) 要充分反映不同的论元角色可能出现的句法位置及其出现条件；

(3) 要充分反映不同的论元角色的形态特征，比如在什么位置上出现时必须带什么样的格标记(case marker)；比如：A＋把 P＋VP(小李把书扔了)，A＋向 Go＋VP (老人向小巷深处走去)；

(4) 要充分利用动词的各种形态标记，比如：L＋V 着＋Th (桌子上放着一部《康熙字典》)→＊L＋A＋V 着＋Th(＊桌子上我放着一部《康熙字典》)～L＋V 了＋Th(桌子上放了一部《康熙字典》)→L＋A＋V 了＋Th(桌子上我放了一部《康熙字典》)。

4 各别论元角色的定义和句法、语义特点

这一部分尝试对现代汉语动词常见的 17 种论元角色，先给出其宽泛的语义定义并辅以一定的例证，再描述其动态的语义特征、给出其比较突出的句法特征，最后列表比较。

1. 施事(agent，简称 A)

施事：自主性动作、行为的发出者。例如：

小王吃了一个馒头　　　　弟弟正看电视呢
妹妹笑了　　　　　　　　他们踢了一会儿足球

这些例子中的主语是施事(agent)，它们共有的、动态的语义特点是：(1) 自立性(independent)，即其所指的事物先于动词所表示的事件独立存在；(2) 使动性(causation)，即其所指的事物施行某个动作、

或造成某种事件或状态。其中，及物动词的施事跟受事或结果相对待，因而是施事的典型成员。

2. 感事(sentient，简称 Se)

感事：非自主的感知性事件的主体。例如：

老王认识李校长　　哥哥喜欢武打片
刘老师太累了　　这孩子又困了

这些例子中的主语是感事，它们共有的语义特点是：(1) 自立性，(2) 感知性(sentience and/or perception)，即其所指的事物在由动词所表示的事件中表现出了某种感知能力。支配感事的动词一定是感觉-心理动词(mental/psychical verbs)，其中及物动词的感事比较接近于施事，形容词的感事比较接近于主事。并且，及物动词的感事是跟对象相对待的。

3. 致事(causer，简称 Cau)

致事：某种致使性事件的引起因素。例如：

老师的夸奖使孩子们很兴奋　父亲严峻的脸色叫我们十分害怕
他的成就令同行羡慕　　这种局势让大家惶恐不安

这些例子中的主语是致事(causer)，它们共有的、动态的语义特点是：(1) 自立性，(2) 使动性，即其所指的事物引发了某种感知性事件；(3) 述谓性(predicative)，即它直接和间接地指陈(denote)一个致使性的(causative)事件，正是这个致使性事件作为原因造成了作为结果的某种感知性事件；比如，"老师的夸奖"直接指陈老师夸奖孩子们这件事，"父亲严峻的脸色"间接指陈父亲摆出了严峻的脸色。[①]

4. 主事(theme，简称 Th)

主事：性质、状态或变化性事件的主体。例如：

小王长了一个疖子　　锅里的水开了
小孩掉沟里了　　村后的桥坍了

这些例子中的主语是主事，它们共有的语义特点是：(1) 自立性，

① 关于"使"字句主语的语义特点，详见袁毓林(2002)。

(2) 变化性(change of state),即其所指的事物的状态在由动词所表示的事件中发生了变化。并且,及物动词的主事是跟系事相对待的。

一般地说,支配施事的动词是动作动词和自主动词,支配主事的动词是非动作动词和非自主动词。支配感事的是心理动词和表示感觉的形容词,可以合称为感知动词(包括形容词);在自主性上,它们介于自主和非自主之间。支配致事的是"使、叫(教)、令、让"等数量极少的致使动词,它们都是非自主动词。在意义上,整个施事、感事、致事、主事可以看作是原型施事(proto-typical agent)的四个典型性渐减的小类。可以列表对照如下:

论元\语义特点	自 立	使 动	感 知	述 谓	变 化
施事	+	+	+	—	—
感事	+	—	+	—	—
致事	+	+	—	+	—
主事	+	—	—	—	+

为了方便和周全,我们把主事当作是主体论元的收容所——凡是不便归入施事、致事和感事的主体论元都放进主事。这样,虽然像"是"、"有"一类动词的主语不一定有变化性的特点,但是我们可以从容地把它们纳入主事。

从句法上看,主体论元共有的特点是能作基础句的主语,但是,其相应的谓语动词在形式和意义上都有一定的差别:施事的谓语是自主动词,能受"不"和"没有"修饰,如"不吃"~"没有吃"、"不走"~"没有走";感事的谓语是感知动词,能受"不"修饰、但一般不受"没有"修饰,如"不认识"~"＊没有认识"、"不困"~"＊没有困";致事的谓语是非自主动词,一般不受"不"和"没有"的修饰,如"＊不/? 没有使孩子们高兴"、"＊不/＊没有叫我们害怕"、"＊不/? 没有令同行羡慕"、"＊不/? 没有让大家惶恐不安";主事的谓语是非自主动词,能受"没有"修饰、一般不受"不"修饰,如"没有掉"~"＊不掉"、"没有醒"~"? 不醒"。可以列表对比如下:

论元\句法特点	作基础句的主语	出现在“不 VP”之前	出现在“没有 VP”之前
施事	+	+	+
感事	+	+	+
致事	+	－	－
主事	+	－	+

从分布上看，施事、感事、致事、主事是不能同现的，即没有对立的价值；因此，也可以把这四个论元角色合成一个不加区分，统称为主体。退一步说，即使区分不清这四个论元角色也不影响论元角色系统的大局。

5. 受事（patient，简称 P）

受事：因施事的行为而受到影响的事物。例如：

老陈吃了一个苹果　　弟弟打了一个茶杯

韩老师批评了小刚　　老兵常常欺负新兵

这些例子中的宾语是受事，其语义特点是：(1) 自立性，(2) 变化性，(3) 受动性（causally affected），即其所指事物承受由动词所表示的动作、行为的影响。受事一定是跟施事相对的，它们共同成为某种类型的及物动词的两个必有论元（obligatory arguments）。

6. 与事（dative，简称 D）

与事：动作、行为的非主动的参与者。例如：

张三给了李四一本词典　　老板对雇员发火

小孙问了老师一个问题　　你向当事人打听一下

在这些例子中，动词的近宾语或介词的宾语是与事，其语义特点是：(1) 自立性，(2) 受动性，参与性（participant），即其所指事物自愿或被迫参与到由动词所表示的动作、行为或事件中去。与事一定是跟施事相对的，对于双宾动词来说，与事还跟受事相对。施事、受事和与事共同成为双宾动词的三个必有论元。

7. 结果（result，简称 R）

结果：由施事的动作、行为造成的结果。例如：

妈妈给我织了一件毛衣　　孩子在桌子上踩了一个脚印
爸爸挖了一个菜窖　　他把窗户纸捅了一个窟窿

这些例子中的宾语是结果，其语义特点是：(1) 变化性，(2) 受动性，(3) 渐成性(incremental)，即其所指事物是在由动词所表示的事件中逐步形成的，这一点正好跟自立性相反。结果一定是跟施事相对的，它们共同成为某种类型的及物动词的两个必有论元。

8. 对象(target，简称 Ta)

对象：感知行为的对象和目标(target)。例如：

爸爸认识刘校长　　小王熟悉广告业务
妹妹喜欢芭蕾舞　　李小明相信通灵术

这些例子中的宾语是对象，其语义特点是：(1) 自立性，(2) 关涉性(concerned)，即其所指表示相应感事所感知的对象和目标等关联物。对象一般是跟感事相对的，它们共同成为某种类型的及物动词(主要是心理动词)的两个必有论元。

9. 系事(relevant，简称 Re)

系事：在事件里跟主事相对的事物。例如：

老赵是仓库保管员　　许先生有三个儿子
这些房子属于地质学院　　小平跑第二棒
我们叫她知心姐姐　　郝海东踢中锋

这些例子中的宾语或远宾语是系事，其语义特点是：(1) 自立性，(2) 类属性(classification/attribute)，即其所指表示相应主事的属性、类型等。

在意义上，整个受事、结果、与事、对象和系事可以看作是原型受事(prototypical patient)五个典型性渐减的小类。可以列表对比如下：

论元\语义特点	受动	变化	自立	渐成	关涉	类属
受事	+	+	+	−	+	−
与事	+	−	+	−	+	−

论元\语义特点	受 动	变化	自立	渐 成	关 涉	类 属
结果	+	+	−	+	+	−
对象	−	−	+	−	+	−
系事	−	−	+	−	−	+

为了方便和周全，我们把系事当作是客体论元的收容所——凡是不便归入受事、与事、结果和对象的客体论元都放进系事。这样，像“是”、“有”、“属于”一类动词的宾语，虽然它们没有变化性的特点，但是我们可以从容地把它们纳入系事。

从句法上看，客体论元共有的句法功能是能作基础句的宾语。其中，受事和与事可以作双宾动词的宾语（即分别作远宾语和近宾语），而结果、对象和系事不能；受事和结果可以作介词“把”的宾语，与事、对象和系事不能。可以列表对比如下：

论元\句法特点	作基础句的宾语	作近宾语	作远宾语	作“把”的宾语
受事	+	−	+	+
与事	+	+	−	−
结果	+	−	−	+
对象	+	−	−	−
系事	+	−	+	−

从分布上看，受事、与事、结果和对象在某些句式中是可以同现的，即具有对立的价值；因此，必须把这四个论元角色加以区分。然后，把论旨角色不太明显的客体论元归入系事。

10. 工具（instrument，简称 I）

工具：动作、行为所凭借的器具。例如：

小王用水果刀切黄瓜 ～ 小王切这把水果刀

爸爸用显微镜看切片 ～ 爸爸正看显微镜呢

这些例子中的宾语或介词宾语是工具，其语义特点是：(1) 自立性，(2) 位移性(movement)，即其所指事物在由动词所表示的事件中可以移动位置。

11. 材料(material,简称 Ma)

材料：动作、行为所用的材料。例如：

姐姐用毛线织了一件上衣 ～ 姐姐正织毛线呢

爷爷用米泔水浇花 ～ 爷爷正浇米泔水呢

这些例子中的宾语或介词宾语是材料，其语义特点是：(1) 自立性，(2) 位移性，(3) 变化性，即其所指的事物在动作、行为中消耗掉了或者由原料变为成品。

12. 方式(manner,简称 M)

方式：动作行为所采取的方式、方法。例如：

他用低音唱了一首《船夫曲》～ 余子真一向唱低音

这些纸包得捆双十字　　这些软糖你还是包小包吧

这些例子中的宾语或介词宾语是方式，其语义特点是：非自立性和附庸性(existence not independent of event)，即其所指状况依附于由动词所表示的动作、行为之上。

工具、材料和方式这三个凭借论元跟动词所表示的动作、行为关系密切，一般直接融入由动词所表示的事件中。这跟以处所为原型的环境论元不同，环境论元一般为动词所表示的事件设定外部的空间条件。也就是说，在跟动词的语义关系方面，环境论元可能比凭借论元更为外围。凭借论元和环境论元作为外围论元，它们共有的语义特征是：既不具有使动性，也不具有受动性。这是它们区别于核心论元的地方，可以列表对比如下：

论元\语义特点	使动	受动	自立	附庸	位移	变化
工具	－	－	＋	－	＋	－
材料	－	－	＋	－	＋	＋
方式	－	－	－	＋	－	－
处所	－	－	＋	－	－	－

从句法上看，工具、材料、方式、处所等论元角色一般都能在基础句中作介词的宾语。其中，工具可以作“用”的宾语，还能通过话题化

而作主语。材料可以作“用”和“把”的宾语，还能通过话题化而作主语。方式可以作“用”的宾语，一般不能话题化。处所可以作“在”的宾语，还能通过话题化而作主语。可以列表对比如下：

论元\句法特点	作介宾	“用”之宾	“把”之宾	“在”之宾	话题化
工具	+	+	−	−	+
材料	+	+	+	−	+
方式	+	+	−	−	−
处所	+	−	−	+	+

上面我们用处所作为代表，来说明环境论元跟凭借论元在句法、语义上的差别。下面，我们把环境论元分为场所、源点和终点三种论旨角色，来具体地讨论一下。

13. 场所(location，简称 L)

场所：动作、行为发生的处所。例如：

小王在里圈跑～小王跑里圈　老刘在食堂吃饭～老刘吃食堂

老侯在江湖上闯荡了几十年 ～ 老侯闯荡江湖几十年

14. 源点(source，简称 So)

源点：动作、行为开始的地点或时间。例如：

一个犯人从监狱里跑了～(从)监狱里跑了一个犯人

一块石头从山顶上滚下来～(从)山顶上滚下来一块石头

长江发源于青藏高原　　这种制度起源于唐朝

他们昨天离开北京去上海　40 年代她脱离了党组织

15. 终点(goal，简称 Go)

终点：动作、行为结束的地点、时间或状态。例如：

他往桌上放了一本书～桌上放了一本书

嫌疑犯跑国外了

我们村来了三个知青　　搞阴谋的人必定以失败而告终

孩子去姥姥家了　　火车正点到达北京站

场所、源点和终点三种论元角色主要跟处所相关，可以总称为处

所论元。这三种论元角色的区别是：源点可以用在“自、从”一类介词之后、不能用在“在、到、向、往”一类介词之后，可以通过话题化而作主语或通过述题化而作宾语；终点可以用在“在、到、向、往”一类介词之后、不能用在“自、从”一类介词之后，可以通过话题化而作主语或通过述题化而作宾语；场所则不论起点和终点，可以用在“在”一类介词之后，不能用在“自、从”或“到、向、往”一类介词之后，不能通过话题化而作主语但可以通过述题化而作宾语。它们在句法上的差别可以列表对比如下：

论元\句法特点	“在”宾	“从”之宾	“往”之宾	话题化	述题化
场所	+	−	−	−	+
源点	−	+	−	+	+
终点	+	−	+	+	+

16. 范围(range，简称 Ra)

范围：动作、行为所涉及的数量、频率、幅度、时间等相关事项。例如：

一个西瓜买三块钱　　一个小时跑二十公里
渔船偏离了主航道几百米　　他老是逃避家务
会议持续了三个小时　　双方僵持了半年
他们休息星期天　　老王值星期六，我值星期天

为了方便和周全，我们把范围当作是外围论元的收容所——凡是不便归入凭借论元和处所论元的其他外围论元都可以放进范围。

17. 命题(proposition，简称 Pn)

命题：由主谓结构、述宾结构或动词、形容词等谓词性成分充当的论元，它本身具有一个由谓词及其论元构成的论元结构；在外部的语义功能上，它以整体充当主体论元或客体论元。例如：

大家认为这事不赖小王　　小刘跳槽影响了达利公司的声誉
他们迫使小刚逃离家乡　　中国女排力争夺取奥运会金牌

5 不同的论元角色之间的配合关系

不同的论元角色之间有着严格的同现限制关系，表现为：有的论元角色可以共现、有的论元角色不能共现、有的论元角色出现时强制性地要求某种论元角色共现。下面，我们通过例子来说明哪些论元角色可以跟哪些论元角色共现。

(1) 施事可以单独出现，或者跟受事、与事、结果等相配对，可以表示为：

S1：A＋__；　例如：老王走了。

S2：A＋__＋P；　例如：爸爸买了一本书。

S3：A＋__＋R；　例如：哥哥做了一张椅子。

S4：A＋__＋D＋P；　例如：宋老师给了我两本词典。

(2) 感事可以单独出现，或者跟对象等相配对，可以表示为：

S5：Se＋__；　例如：爷爷困了。

S6：Se＋__＋Ta；　例如：老陈认识许主任。

(3) 主事可以单独出现，或者跟系事等相配对，可以表示为：

S7：Th＋__；　例如：炉膛里的火灭了。

S8：Th＋__＋Re；　例如：叶文龙是一条好汉。

(4) 工具、材料、方式等一般跟施事及其配对成分受事和结果共现，可以表示为：

S9：A＋用I＋__＋P；　例如：妈妈用小刀切西瓜。

S10：A＋用I＋__＋R；　例如：爸爸用铁锹挖了一个菜窖。

S11：A＋用Ma＋__＋P；　例如：奶奶用米泔水浇兰花。

S12：A＋用Ma＋__＋R；　例如：姐姐用细毛线织了一件上衣。

S13：A＋用M＋__＋P；　例如：老张用双十字捆被子。

S14：A＋用M＋__＋R；　例如：孙晓平用高音唱了一首歌。

S15：A＋__＋I；　例如：我切这把大刀。

S16：A＋__＋Ma；　例如：妈妈正在织毛线呢。

S17：A＋__＋M；　例如：我捆双十字。

(5) 致事一般跟感事配对出现，可以表示为：

S18：Cau＋__＋Se＋VP；例如：爷爷的身体使全家人担心。

6 结语：论元角色变化的语义机制

显然，上文讨论的17种论元角色肯定不能涵盖现代汉语动词的论元的所有的论旨角色。为了有效地说明句子的结构形式跟意义结构之间的关系，分化和说明歧义句式，上文所列的论元角色在必要时须作更为精细的分别。比如，施事有时须分为施益性的和受益性的，与事有时须分为目标性(或受益性)的和来源性(或施益性)的，受事有时须分为对象性的和范围性的，结果有时须分为同源性的、后果性的和成品性的，工具有时须分为人体性的、器具性的和材料性的。例如：

(1) a. 小王正(给他孩子)理发呢

b. 小王正(在理发店)理发呢

c. 小王正在理发呢

(2) a. 刘为借给老张一辆自行车

b. 刘为向老张借一辆自行车

c. 刘为借老张一辆自行车

(3) a. 张老三正浇草坪呢

b. 张老三正浇菜苗呢

c. 张老三正浇菜园呢

(4) a. 这孩子又摔了一个跟斗

b. 这孩子又摔了一个大包

c. 这孩子又画了一幅年画

(5) a. 他用双手贴标语

b. 他用刷子贴标语

c. 他用糨糊贴标语

(1a)中的“小王”是施益性的施事，(1b)中的“小王”是受益性的施事，(1c)中的“小王”是两可的。(2a)中的“老张”是目标性(或受益性)的

与事，(2b)中的“老张”是来源性(或施益性)的与事，(2c)中的“老张”是两可的；(3a)中的“草坪”是范围性的受事，(3b)中的“菜苗”是对象性的受事，(3c)中的“菜园”似乎是两可的。

有时，两种论元角色会合并为一种论元角色，结果可能产生一种新的论元角色。例如：

(6) a. 老王用锤子把屋顶砸了一个洞

b. 陨石把屋顶砸了一个洞

(6b)的“陨石”是动力(force)，相当于(5a)中施事“老王”和工具“锤子”的合并。

有时，不同的论元角色之间还会发生动态语义的转变，从而从一种论元角色转变为另一种论元角色。例如：

(7) a. 他们正用水泵抽着污水呢

b. ？他们正抽着污水呢

c. 水泵正抽着污水呢

(8) a. 我们常常在食堂吃中饭

b. 中饭我们常常吃食堂

(9) a. 他用小棍儿掏鸟窝

b. 他用小棍儿掏鸟蛋

c. 他用小棍儿从鸟窝里掏鸟蛋

(10) a. 爸爸拍了一下我的肩膀

b. 爸爸在我的肩膀上拍了一下

原来(7a)中的工具“水泵”在(7c)中施事化了，原来(8a)中的场所“食堂”在(8b)中受事化了，原来(9a、10a)中的范围性受事“鸟窝”和“我的肩膀”分别在(9c、10b)中处所化了。可见，论元成分的施事化和受事化是以占据主语或宾语位置为语法形式标志的，论元成分的处所化是以后加方位词和占据状语位置为语法形式标志的。

我们相信，引入论元角色的细分、合并和转化等语义机制，[①]是

---

① 详见袁毓林(1998)，第122—142页。

简化描述论元成分的论旨角色的有效手段；否则，设立再多的论旨角色也是难以穷尽所有的论元成分的各种微妙复杂的语义作用的。

参考文献

陈　平（1994）《试论汉语中三种句子成分与语义成分的配位原则》，《中国语文》第3期。

程　工（1995）《评〈题元原型角色与论元选择〉》，《国外语言学》第3期。

顾　阳（1994）《论元结构理论介绍》，《国外语言学》第1期。

韩万衡（1997）《德国配价论主要学派在基本问题上的观点和分歧》，《国外语言学》第3期。

李　洁（1987）《德语配价理论的发展及成就》，《外语教学与研究》第1期。

鲁　川、林杏光（1989）《现代汉语语法的格关系》，《汉语学习》第5期。

孟　琮等（1987）《动词用法词典》，上海辞书出版社。

汤廷池、张淑敏（1996）《论旨网格、原参语法与机器翻译》，《中国语文》第4期。

徐烈炯（1988）《生成语法理论》，上海外语教育出版社。

徐烈炯（1990）《语义学》，语文出版社。

杨成凯（1986）《Fillmore 的格语法理论》，《国外语言学》第1、2、3期。

袁毓林（1998）《汉语动词的配价研究》，江西教育出版社。

袁毓林（2002）《汉语句子的文意不足和结构省略》，《汉语学习》第3期。

Abraham, W. (ed.) (1978) *Valence, Semantic Case and Grammatical Relation*, John Benjamin's.

Dowty, D. (1991) Thematic Proto-Role and Argument Selection. *Language*, Vol. 67, No. 3.

Fillmore, C. (1968) The Case for Case. *Universals in Linguistic Theory*, (ed.) By Emmon Bach and Robert T. Harms, 1—90. New York: Holt, Rinehart & Winston.《"格"辨》，胡明扬译，《语言学译丛》第二辑，第1—117页，中国社会科学出版社，1980年。

Fillmore, C. (1977a) The Case for Case Reopened, in P. Cole & J. M. Sadock (eds.) *Syntax and Semantics*, Vol. 8, *Grammatical Relations*, pp. 59—81. Academic Press.

Fillmore, C. (1977b) Topics in Lexical Semantics, in R. W. Cole (ed.) *Current Issues in Linguistic Theory*, pp. 76—138.

Grimshaw, J. (1990) *Argument Structure*. MIT Press.

Gruber, J. (1976) *Lexical Structures in Syntax and Semantics*. North-Holland.

Hale, K. & S. J. Keyser (1991) *On the Syntax of Argument Structure*. MIT Press.

Jackendoff, R. (1990) *Semantic Structure*. The MIT Press.

Steinberg, D. & Jakobovits, L. (1971) *Semantics: An Interdisciplinary Reader in Philosophy, Linguistics and Psychology*. Cambridge University Press.

Williams, E. (1994) *Thematic Structure in Syntax*. MIT Press.

2001年4月初稿,2002年2月改定

(发表于《世界汉语教学》2002年第3期)

一套汉语动词的论元角色的语法指标

本文考察了现代汉语动词常见的17种论元角色的语法表现，并以这些论元角色的下列分布特征和转换特征作为测试条件：(i) 能否直接作主语或宾语等句法成分，(ii) 能否作"把/被"字句的主语，(iii) 能否作介词"把/被/由"的宾语，(iv) 能否作"用/在/到/从/往/向"等介词的宾语，(v) 能否作"VV/V一V/V了V"等动词重叠形式的主语或宾语，(vi) 能否作"不/没有VP"等否定形式的主语或宾语，(vii) 能否作"V-成"一类复合动词的宾语，(viii) 能否通过左向或右向出位而成为话题或述题等，从语法形式上界定不同的论元角色应该具备的一套语法指标，希望为定义不同的论元角色提供可把握的操作程序。

1 引 言

动词的论旨角色(thematic role)是根据不同的论元(argument)跟动词的语义关系而划分出来的，或者说是根据论元在由动词及其论元构成的述谓结构(predication)中的语义作用而确立的。比如，如果一个名词性成分在述谓结构中具有施动性，那么这个论元的论旨角色(简称"论元角色")就是施事；如果一个名词性成分在述谓结构中具有受动性，那么这个论元的论旨角色就是受事。显然，论元角色是根据论元成分在述谓结构中动态的语义特征而划分出来的。由于这种语义特征具有较大的模糊性，因而在具体的处理上难免见仁见智；表现为：(1) 不同的学者所设立的论元角色的数目可能相当悬殊，(2) 对不同的论元角色的定义相差极大、并且彼此之间可能难以对应和折合。这种局面导致在论元角色的系统上缺少可比性，使从事汉语信息处理和汉语教学等实际应用工作的人士无所适从。

有鉴于此，我们考察了现代汉语中动词常见的17种论元角色的

语法表现(grammatical behaviour),着重以下列句法分布特点或句法转换特点为测试条件:

(i) 能否直接作主语或宾语等句法成分,

(ii) 能否作"把/被"字句的主语,

(iii) 能否作介词"把/被/由"的宾语,

(iv) 能否作"用/在/到/从/往/向"等介词的宾语,

(v) 能否作"VV/V一V/V了V"等动词重叠形式的主语或宾语,

(vi) 能否作"不/没有VP"等否定形式的主语或宾语,

(vii) 能否作"V-成"一类复合动词(或动词结构)的宾语,

(viii) 能否通过左向或右向出位而成为话题或述题,等等。

以此作为从语法形式上界定不同的论元角色的一套语法指标(grammatical guidelines)。

2 各种论元角色的语法指标

下面依次描述17种论元角色的主要的语法表现,希望它们能够成为界定不同的论元角色的一整套具有一定的可操作性的语法指标。[①]

1. 施事

(1) 作基础句的主语,例如:

弟弟哭了|小张吃了一碗汤面

(2) 作"把"字句的主语,例如:

小张把那碗米饭吃了|哥哥把校长给得罪了

(3) 作"被、由"等介词的宾语,例如:

那碗米饭被小张吃了|后勤工作由老刘负责

(4) 作"V(一)V、V(了)V"等重叠形式的主语,例如:

你试(一)试|妈妈笑(了)笑|李二婶撇了撇嘴

① 本文在描述论元角色的语法表现时,主要以朱德熙(1982)为参照系统。

(5) 作"不 VP"和"没有 VP"等否定形式的主语,例如:

小沈不去～小沈没有去

刘伟不考研究生～刘伟没有考研究生

(6) 一般不在宾语位置上出现,除非句首是处所性成分,例如:

小明笑了～*笑了小明|客人来了～家里来客人了

(7) 作主语时可以左向出位(left-dislocation)而成为话题,原来的位置上可以是句法空位(syntactic gap),留下的空位也可由续指代词(resumptive pronoun)"他"等来填充,例如:

我妹妹,〔她〕笑个不停|小刚,〔他〕考上了清华大学

2. 感事

(1) 作基础句的主语,例如:

弟弟累了|小张认识王大夫|哥哥非常喜欢西洋画

(2) 不作"把"字句的主语,例如:

*小张把王大夫认识了|*哥哥把西洋画非常喜欢

(3) 不作"被、由"等介词的宾语,例如:

*王大夫被小张认识了|*西洋画由哥哥喜欢

(4) 不作"V(一)V、V(了)V"等重叠形式的主语,例如:

*你累(一)累|*妈妈困(了)困

*李二婶认识了认识王老师

(5) 作"不 VP"这种否定形式的主语,不作"没有 VP"这种否定形式的主语,例如:

小沈不累～*小沈没有累

刘伟不认识厂长～*刘伟没有认识厂长

(6) 不在宾语位置上出现,即由它造成的主谓结构不能直接转换为述宾结构,例如:

小明累了～*累了小明|客人困了～*困了客人

(7) 作主语时可以左向出位而成为话题,原来的位置上可以是句法空位,留下的空位可由续指代词"他"等来填充,例如:

我妹妹,〔她〕累得直不起腰|

小刚,〔他〕认识我们学校的陈会计

3. 致事

(1) 作基础句的主语，例如：

弟弟的话使我很难堪|小张的处境叫大家担心

哥哥的到来让奶奶感到突然|公司破产的消息令股民失望

(2) 不作“把”字句的主语，例如：

* 弟弟的话把我很难堪

* 小张的处境把大家担心坏了

* 哥哥的到来把奶奶感到突然

* 公司破产的消息把股民失望极了

(3) 不作“被、由”等介词的宾语，例如：

* 我被弟弟的话很难堪

* 大家被小张的处境担心坏了

* 奶奶由哥哥的到来感到突然

* 股民由公司破产的消息失望极了

(4) 不作“V(一)V、V(了)V”等重叠形式的主语，例如：

* 弟弟的话使(一)使我很难堪

* 小张的处境叫(一)叫大家担心

* 哥哥的到来让了让奶奶感到突然

* 公司破产的消息令了令股民失望

(5) 一般不作“不VP”和“没有VP”等否定形式的主语，例如：

* 弟弟的话不使我很难堪

? 小张的处境没有叫大家担心

* 哥哥的到来不让奶奶感到突然

? 公司破产的消息没有令股民失望

(6) 不能在宾语位置上出现，即由它造成的主谓结构不能直接转换为述宾结构，例如：

弟弟的话使我很难堪～* 我使弟弟的话很难堪

小张的处境叫大家担心～* 大家叫小张的处境担心

(7) 作主语时可以左向出位而成为话题，原来的位置上只能是句法空位，即留下的空位上不能用续指代词“他、这”等来填充，例如：

弟弟的话,〔*这〕使我很难堪

小张的处境,〔*那〕叫大家担心

哥哥的到来,〔*这事〕让奶奶感到突然

公司破产的消息,〔*那〕令股民失望

4. 主事

(1) 作基础句的主语,例如:

弟弟长了一个疖|小张掉了一个钱包

河里的冰都化了|墙也坍了

(2) 一般不作"把"字句的主语,例如:

?小张把钱包丢了|?? 哥哥把手表掉了

*弟弟把疖长在手指上

(3) 一般不作"被、由"等介词的宾语,例如:

?我的书被小张丢了|*那块手表由哥哥掉河里了

(4) 不作"V(一)V、V(了)V"等重叠形式的主语,例如:

*你丢(一)丢钱包|*妈妈掉(了)掉手表

*弟弟长了长疖子

(5) 不作"不VP"这种否定形式的主语,可作"没有VP"这种否定形式的主语,例如:

*小沈不丢钱包～小沈没有丢钱包

*刘伟不掉东西～刘伟没有掉东西

(6) 一般不在宾语位置上出现,除非句首是处所性成分,例如:

河里的冰都化了～*化了河里的冰～?河里化了不少冰

东面的墙都坍了～*坍了东面的墙～牲口棚上坍了一面墙

(7) 作主语时一般不能左向出位而成为话题,例如:

?我妹妹,〔她〕又丢了一个钱包

*小刚,〔他〕昨天掉了一个钱包

5. 受事

(1) 作基础句的宾语,作双宾语句中的远宾语(直接宾语),例如:

弟弟吃了一个苹果|李院长批评了王大夫

妈妈给弟弟一个皮球|老张抽了我一包万宝路

(2) 作“把”等介词的宾语，不作“为、对、给、向、替”等介词的宾语，例如：

弟弟把那个苹果吃了|李院长把王大夫批评了一顿

妈妈把皮球给了弟弟|＊李院长为王大夫批评了一顿

＊弟弟对那个苹果吃了|＊李院长给王大夫批评了一顿[①]

＊妈妈向皮球给了弟弟|＊李院长替王大夫批评了一顿

(3) 作“被”字句的主语，例如：

苹果被弟弟吃了|王大夫被李院长批评了一顿

那个皮球被妈妈给了邻居家的孩子

(4) 作“V(一)V、V(了)V”等重叠形式的宾语，例如：

点(一)点人数|汇报(了)汇报情况|催(一)催王老师

(5) 作“不VP”和“没有VP”等否定形式的宾语，例如：

不了解情况～没有了解情况|不吃米饭～没有吃米饭

(6) 不作基础句的主语，即由它造成的述宾结构不能转换为相应的主谓结构，例如：

了解案情～＊案情了解|接待客人～＊客人接待

吃馒头～＊馒头吃

(7) 作动词的宾语时可以左向出位而成为话题，原来的位置上可以是空位，留下的空位也可由续指代词“他”等形式填充；作介词“把”的宾语时可以左向出位而成为话题，原来的位置上不可以是空位，这个空位必须用续指代词“他”等形式填充。例如：

大闸蟹，我吃过〔这种东西〕|小明，我见过〔他〕

那支毛笔，我已经把它扔了|我的词典，你把它搁哪儿了

(8) 不作“V-成”一类复合动词的宾语，例如：

＊吃成馒头|＊批评成小李

---

① 在“李院长给王大夫批评了一顿”中，当“王大夫”是施事时，这个句子是合格的；这时，介词“给”用在受事主语句里引导施事，其作用跟“叫、让、被”相似；并且，施事可以省略，说成：“李院长给批评了一顿”。参考朱德熙(1982)，第179—180页。

6. 与事

(1) 一般不作基础句的宾语,可作双宾语句中的近宾语(间接宾语),例如:

弟弟借孙老师一本词典|老张问我一个问题

(2) 不作"把"等介词的宾语,作"为、对、给、向、替"等介词的宾语,例如:

大家一起为子孙后代造福|老板对雇员发火

医生给病人把脉|弟弟向目击者打听过这事

我替你把关|*老张把我问一个问题

(3) 不作"被"字句的主语,例如:

*孙老师被弟弟借了一本词典|*病人被王大夫把脉

*邻居家的孩子被我妈妈给了一个皮球

(4) 一般不作"V(一)V、V(了)V"等重叠形式的宾语,例如:

?问(一)问厂长房子的事|*送(了)送王老师一束花

(5) 可作"不 VP"和"没有 VP"等否定形式的宾语,例如:

不问老刘~没有问老刘|不送他们大米~没有送他们大米

(6) 不作基础句的主语,即由它造成的述宾结构不能转换为相应的主谓结构,例如:

问小刘(一件事)~*小刘问(一件事)

送他(一本书)~*他送(一本书)

(7) 作动词的宾语时可以左向出位而成为话题,留下的空位必须用续指代词"他"等填充;例如:

陈先生,小方问过他股票行情

结婚的,我都送他们一套炊具

(8) 不作"V-成"一类复合动词的宾语,例如:

*问成小刘一件事|*送成小李一本书

7. 结果

(1) 作基础句的宾语,不作双宾语句中的宾语,例如:

爸爸在院子里挖了一口井|小王在桌上踩了一个脚印

妈妈给弟弟织了一件毛衣|老张在窗户纸上捅了一个洞

(2) 作"把"等介词的宾语,不作"为、对、给、向、替"等介词的宾

语，例如：

弟弟把纸船叠好了|李院长把通知写黑板上了

他们把房子盖在山坡上了|＊李院长为通知写黑板上了

＊弟弟对纸船叠好了|＊他们给房子盖在山坡上了

＊李院长向通知写黑板上了|＊弟弟替纸船叠好了

(3) 作“被”字句的主语，例如：

房子被他们盖好了|便桥被工兵一夜之间架起来了

菜窖被爸爸挖在院子里

(4) 作“VV、V — V”等重叠形式的宾语，不作“V 了 V”等重叠形式的宾语，例如：

做(一)做饭|写(一)写信|＊做了做饭|＊写了写信

(5) 作“不 VP”和“没有 VP”等否定形式的宾语，例如：

不造桥～没有造桥|不做米饭～没有做米饭

(6) 不作基础句的主语，即由它造成的述宾结构不能转换为相应的主谓结构，例如：

烧米饭～＊米饭烧|制造谣言～＊谣言制造

捏饺子～＊饺子捏

(7) 作动词的宾语时可以左向出位而成为话题，留下的空位上不能填入续指代词“他”等形式；作介词“把”的宾语时可以左向出位而成为话题，原来的位置上不可以是空位，这个空位必须用续指代词“他”等填充；例如：

毛衣，妈妈早就为我织好〔＊它〕了

菜窖，爸爸已经挖好〔＊它〕了

招工广告，我把它写在大门口了

百科词典，我们把它编出来了

(8) 作“V-成”一类复合动词的宾语，例如：

揉成馒头|做成工棚|挖成深井|盖成电脑超市

8. 对象

(1) 作基础句的宾语，作双宾语句中的近宾语(间接宾语)；例如：

弟弟喜欢西洋美术|厂长欣赏他的才能

我认识他们老板|校长非常信任李晓明

我们喜欢他勤奋踏实|大伙儿讨厌他太啰唆

(2) 不作"把"等介词的宾语，也不作"为、对、给、向、替"等介词的宾语，例如：

*哥哥把漫画喜欢得不得了|*妈妈把这些人讨厌透了

*爸爸把李校长认识了|*张书记把李晓明信任极了

*哥哥为漫画喜欢得不得了|*妈妈对这些人讨厌透了

*爸爸给李校长认识了|*张书记向李晓明信任极了

*哥哥替漫画喜欢得不得了

(3) 不作"被"字句的主语，例如：

老陈熟悉财会工作～*财会工作被老陈熟悉

我相信功能主义～*功能主义被我相信

哥哥认识小方～*小方被哥哥认识

(4) 不作"V(一)V、V(了)V"等重叠形式的宾语，例如：

*喜欢(一)喜欢评弹|*讨厌(了)讨厌说教

*信(一)信西医

(5) 作"不 VP"这种否定形式的宾语，不作"没有 VP"这种否定形式的宾语，例如：

不相信群众～*没有相信群众

不喜欢中餐～*没有喜欢中餐

(6) 不作基础句的主语，即由它造成的述宾结构不能转换为相应的主谓结构，例如：

喜欢大山～*大山喜欢|讨厌礼节～*礼节讨厌

认识你爸～*你爸认识

(7) 作动词的宾语时可以左向出位而成为话题，原来的位置上可以是空位，留下的空位也可由续指代词"他"等形式填充；例如：

西洋画，我喜欢〔这种东西〕|小明，我讨厌〔他〕

那个孩子，我认识〔她〕|这些家伙，我熟悉〔他们〕

(8) 不作"V-成"一类复合动词的宾语，例如：

*认识成小李|*喜欢成漫画

9. 系事

(1) 作基础句的宾语,有时作双宾语句中的远宾语(直接宾语);例如:

弟弟是北大附中的学生|这些房子属于地质学院

我们都叫他华威先生|工人们骂他狗奴才

(2) 不作"把"等介词的宾语,也不作"为、对、给、向、替"等介词的宾语,例如:

哥哥是工人~*哥哥把工人是

这些房子属于地质学院~*这些房子为地质学院属于

知青们叫我妈大姐~*知青们把/对/给/向/替大姐叫我妈

(3) 不作"被"字句的主语,例如:

苹果属于水果~*水果被苹果属于

小平姓尹~*尹被小平姓

老陈是工会主席~*工会主席被老陈是

(4) 不作"VV、V一V、V了V"等重叠形式的宾语,例如:

*像(一)像他爸|*拥有(了)拥有财产|*是(一)是师傅

(5) 作"不VP"这种否定形式的宾语,不作"没有VP"这种否定形式的宾语,例如:

不是学生~*没有是学生|不属于国家~*没有属于国家

(6) 不作基础句的主语,即由它造成的述宾结构不能转换为相应的主谓结构,例如:

像大山~*大山像|有力气~*力气有|是工人~*工人是

(7) 作动词的宾语时一般不能左向出位而成为话题,只有作"有、不是"等的宾语时才可以左向出位而成为话题、留下的空位不能由续指代词"他"等填充;例如:

*大胖熊,你简直像〔 〕|*集体,荣誉属于〔 〕

*狂热分子,我是〔 〕~狂热分子,我不是〔 〕|毛笔,我有〔 〕

(8) 不作"V-成"一类复合动词的宾语,例如:

*是成厂长|*属于成校方

10. 工具

(1) 作介词"用"的宾语,整个介宾结构在基础句中放在动词之

前作状语，例如：

爸爸用小刀切萝卜|爷爷用放大镜看报纸

(2) 不作介词“把”的宾语，例如：

*爸爸把小刀切萝卜|*爷爷把放大镜看报纸

(3) 不作介词“在、从、到”等的宾语，例如：

*爸爸在小刀切萝卜|*爷爷从放大镜看报纸

(4) 可以左向出位而成为话题，留下的空位必须用续指代词“他”等填充，或者把介词“用”删除，例如：

这把斧子，叔叔用它砍柴|这副眼镜，我看电视

(5) 可以右向出位(right dislocation)而成为述题，即紧接在动词之后作宾语，留在原位的介词“用”必须删除，例如：

叔叔砍这把斧子|你看这个望远镜|我切水果刀

(6) 可作“VV”等重叠形式的宾语，不作“V 一 V、V 了 V”等重叠形式的宾语，例如：

你切切这把刀试试|我也砍砍这把斧子

*看一看这副望远镜|*看了看那副眼镜

*切一切那把水果刀|*砍了砍这把斧子

(7) 不作“V-成”一类复合动词的宾语，例如：

*切成水果刀|*看成那副眼镜

11. 材料

(1) 作介词“用”的宾语，整个介宾结构在基础句中放在动词之前作状语，例如：

爸爸用柳条编箱子|爷爷用米泔水浇花|我用白灰刷墙

(2) 作介词“把”的宾语，例如：

爸爸把柳条都编了箱子了|妈妈把米泔水都浇了花了

叔叔把白灰全刷墙上了

(3) 不作介词“在、从、到”等的宾语，例如：

*爸爸在柳条都编了箱子了|*妈妈从米泔水都浇了花了

*叔叔到白灰全刷墙上了

(4) 可以左向出位而成为话题，留下的空位必须用续指代词“他”等填充，或者把介词“用”删除，例如：

这些柳条，叔叔用它编花篮|这些米泔水，我浇兰花

(5) 可以右向出位而成为述题，即紧接在动词之后作宾语，留在原位的介词“用”必须删除，例如：

叔叔正编柳条呢|这盆兰花我浇米泔水|我爸正刷石灰水呢

(6) 可作“VV”等重叠形式的宾语，不作“V一V、V了V”等重叠形式的宾语，例如：

晚上没事就编编柳条|你来刷刷这种石灰水

*编一编柳条|*浇了浇米泔水|*编了编柳条

(7) 不作“V-成”一类复合动词的宾语，例如：

*编成柳条|*浇成米泔水

12. 方式

(1) 作介词“用”的宾语，整个介宾结构在基础句中放在动词之前作状语，例如：

爸爸用低音唱了一首民歌|爷爷用三角包包糖果

王蒙用意识流写小说|她用花腔唱《茶花女》

(2) 不作介词“把”的宾语，例如：

*爸爸把低音唱了一首民歌|*爷爷把三角包包糖果

*王蒙把意识流写小说|*她把花腔唱《茶花女》

(3) 不作介词“在、从、到”等的宾语，例如：

*爸爸在低音唱了一首民歌|*爷爷在三角包包糖果

*王蒙到意识流写小说|*她从花腔唱《茶花女》

(4) 不能左向出位而成为话题，例如：

*低音，爸爸用它唱了一首民歌

*三角包，爷爷用它包糖果

*意识流，王蒙用它写小说

*花腔，王芳用它唱《茶花女》

(5) 可以右向出位而成为述题，即紧接在动词之后作宾语，留在原位的介词“用”必须删除，例如：

刘刚唱低音，张平唱高音|爷爷正包三角包呢

他尝试写意识流|这条被子，我捆井字

(6) 不作“VV、V一V、V了V”等重叠形式的宾语，例如：

*晚上没事就唱唱低音|? 你来包包这种三角包|

*唱一唱花腔|*写一写意识流

*唱了唱高音|*捆了捆双十字

(7) 可以作“V-成”一类复合动词的宾语，例如：

唱成高音了|包成三角包|捆成井字

13. 场所

(1) 作介词“在”的宾语，整个介宾结构在基础句中放在动词之前作状语，例如：

黄斌在里圈跑|爷爷在食堂吃中饭

老侯在江湖上闯荡了几十年

老陈在地板上睡|张书记老在乡下住

(2) 不作介词“把、用”等的宾语，例如：

*黄斌把里圈跑|*爷爷把食堂(里)吃中饭

*老陈用地板上睡|*张书记老用乡下住

(3) 不作介词“从、往、向”等的宾语，例如：

*黄斌从里圈跑|*爷爷从食堂(里)吃中饭

*老陈往地板上睡|*张书记老往乡下住

(4) 不能左向出位而成为话题，例如：

*里圈，黄斌在那儿跑|*食堂(里)，爷爷在那儿吃中饭

*地板上，老陈在那儿睡|*乡下，张书记老在那儿住

(5) 可以右向出位而成为述题，即紧接在动词之后作宾语；这时，附着在处所性成分之后的方位词必须删除，留在原位的介词也必须删除；例如：

黄斌跑里圈，刘虹跑外圈|爷爷老吃食堂

老陈经常睡地板|张书记老住乡下|老侯闯荡江湖几十年

(6) 作“VV、V一V”等重叠形式的宾语，不作“V了V”等重叠形式的宾语，例如：

你跑(一)跑外圈看|你吃(一)吃食堂看

*老陈睡了睡地板|*黄斌跑了跑里圈

*张书记住了住乡下|*我也吃了吃食堂

(7) 不作“V-成”一类复合动词的宾语，例如：

＊跑成里圈｜＊吃成食堂

14. 源点

(1) 作介词“从”的宾语，整个介宾结构在基础句中放在动词之前作状语；或者作介词“于”的宾语，整个介宾结构在基础句中放在动词之后作补语。例如：

一个犯人从监狱里跑了｜一只鹦鹉从笼子里飞了

长江发源于青藏高原｜科举制度起源于隋唐

(2) “离开”等极少数动词的源点只能作宾语，例如：

父亲十八岁就离开了故乡｜代表团离开北京去广州参观

(3) 不作介词“把、用”等的宾语，例如：

＊一个犯人把监狱里跑了｜＊一只鹦鹉用笼子里飞了

＊长江发源把青藏高原｜＊科举制度起源用隋唐

(4) 不作介词“在、往、到”等的宾语，例如：

＊一个犯人在监狱里跑了｜＊一只鹦鹉往笼子里飞了

＊长江发源在青藏高原｜＊科举制度起源往隋唐

(5) 用介词“从”引导的源点可以左向出位而成为话题，留在原位的介词“从”必须删除；同时，跟源点相应的施事或主事必须通过述题化而移到动词之后作宾语；用介词“于”引导的源点不能左向出位而成为话题；例如：

监狱里跑了一个犯人｜笼子里飞了一只鹦鹉

＊青藏高原，长江发源于此｜＊隋唐，科举制度起源于那时

(6) 不能右向出位而成为述题，例如：

＊一个犯人跑了监狱(里)｜＊一只鹦鹉飞了笼子(里)

＊长江发源青藏高原｜＊科举制度起源隋唐

(7) 不作“VV、V一V、V了V”等重叠形式的宾语，例如：

＊跑(一)跑监狱｜＊飞(一)飞笼子｜＊跑了跑监狱

＊飞了飞笼子｜＊发源发源青藏高原｜＊起源了起源唐朝

(8) 不作“V-成”一类复合动词的宾语，例如：

＊跑成监狱里｜＊起源成唐朝

15. 终点

(1) 作介词“向、往”等的宾语，整个介宾结构在基础句中放在动

词之前作状语；或者作介词“到、向、往、在”等的宾语，整个介宾结构在基础句中放在动词之后作补语；例如：

他们向科学高峰攀登｜她往书架上插书

这趟车开往齐齐哈尔｜刘磊把书放在家里了

(2) “去、到、到达”等少数动词的终点只能作宾语，例如：

小王去图书馆了｜孩子到姥姥家了｜他们明天到达深圳

(3) 不作介词“把”、“用”等的宾语，例如：

*他们把科学高峰攀登｜*她用书架上插书

(4) 不作介词“从”等的宾语，例如：

*他们从科学高峰攀登｜*她从书架上插书

(5) 能用介词“在”引导的终点可以左向出位而成为话题，留在原位的介词“在”必须删除；能用介词“到”引导的终点可以左向出位而成为话题，留下原位的介词“到”必须删除，同时跟终点相应的施事必须通过述题化而移到动词之后作宾语；用介词“往、向、到”等介词引导的终点，不能左向出位而成为话题。例如：

小明在桌上放了一本书～桌上小明放了一本书

两个乡干部来到我们村～我们村来了两个乡干部

这趟车开往齐齐哈尔～*齐齐哈尔，这趟车开往〔那儿〕

他们向科学高峰攀登～*科学高峰，他们攀登〔那儿〕

(6) 可以右向出位而成为述题，例如：

他们勇敢地攀登科学高峰｜那本书，她插书架上了

这趟车开齐齐哈尔｜老刘把资料搁家里了

(7) 不作“VV、V一V、V了V”等重叠形式的宾语，例如：

*插(一)插书架(上)｜*开(一)开齐齐哈尔

*搁了搁家里｜*坐了坐沙发(上)

*攀登了攀登科学高峰｜*放了放桌子(上)

(8) 不作“V-成”一类复合动词的宾语，例如：

*攀登成科学高峰｜*插成书架上｜*开成齐齐哈尔

*搁成家里

16. 范围

(1) 作基础句的宾语，有时作双宾语句中的远宾语，例如：

这把椅子卖两百块钱|那套房子值四五十万

这辆车一天跑了几百里|炸弹偏离目标三十米

会议持续了三个小时|双方僵持了半年

(2) “休息、值(班)”等极少数动词的表示时间的范围论元，通常在介词“在”的引导下作状语，但也可以右向出位通过述题化而成为宾语，例如：

他们在星期天休息 ～他们休息星期天

老王在星期六值(班)，我在星期天值(班)～

老王值星期六，我值星期天

(3) 不作“把、用”等介词的宾语，不作“为、对、给、向、替”等介词的宾语，也不作介词“在、从、到、往”等的宾语；例如：

＊这把椅子把两百块钱买|＊那套房子用四五十万值

＊这辆车一天为几百里跑

＊炸弹从/往/在/到三十米偏离目标

(4) 不作“被”字句的主语，例如：

＊两百块钱被这把椅子买|＊四五十万被那套房子值

＊几百里被这辆车一天跑|＊三十米被炸弹偏离目标

(5) 不作“被、由”等介词的宾语，例如：

＊这把椅子被两百块钱买|＊那套房子被四五十万值

＊这辆车一天由几百里跑|＊炸弹由三十米偏离目标

(6) 不作“VV、V一V、V了V”等重叠形式的宾语，例如：

＊买(一)买两百块钱|＊值(一)值四五十万

＊跑了跑几百里|＊偏离了偏离三十米

(7) 一般不能左向出位而成为话题，例如：

＊两百块钱，这把椅子买|？四五十万，那套房子值

＊八百里，这辆车一天跑了|＊三十米，炸弹偏离目标

(8) 不作“V-成”一类复合动词的宾语，例如：

＊买成两百块钱|＊值成四十万|＊跑成几百里

＊偏离成三十米

17. 命题

(1) 作基础句的主语或宾语,作主语时可以用代词"这(事)、那(事)"等来称代和替换,作宾语或宾语补足语时可用"这样(做)、那样(做)"等形式来称代和替换。例如:

老王旷工影响了生产进度～这(事)影响了生产进度

刘老师带病上课感动了学生～那(事)感动了学生

他们觉得那地方不错～他们觉得这样/这样觉得

大伙儿认为我没理～大伙儿认为这样/这样认为

他们企图从背后下手～他们企图那样(做)

弟弟打算盖一座楼房～弟弟打算那样(做)

他们迫使我放弃学位～他们迫使我这样(做)

(2) 不作"把、用"等介词的宾语,不作"为、对、给、向、替"等介词的宾语,也不作介词"在、从、到、往"等的宾语,例如:

*他们把那个地方不错觉得|*大伙儿用我没理认为

*他们为从背后下手企图

*弟弟从/往/在/到盖一座楼房打算

(3) 不作"被"字句的主语,例如:

*那个地方不错被他们觉得|*我没理被大伙儿认为

*从背后下手被他们企图|*盖一座楼房被弟弟打算

(4) 不作"被、由"等介词的宾语,例如:

*生产进度被老王旷工影响了

*生产进度由老王旷工影响了

*学生被刘老师带病上课感动了

*学生由老师带病上课感动了

(5) 不作"VV、V一V、V了V"等重叠形式的宾语,例如:

*觉得(一)觉得那地方不错|*认为(一)认为我没理

*企图了企图从背后下手|*打算了打算盖一座楼房

(6) 作主语时可以左向出位而成为话题,留下的空位可以用续指代词"这/那(事)"等填充;作宾语时不能左向出位而成为话题,但可以通过易位而前置到主语之前;例如:

老王旷工,〔这(事)〕影响了生产进度

刘老师带病上课,〔那(事)〕感动了学生

那地方不错'他们觉得～＊那地方不错,他们觉得〔这样〕[1]

我没理'大伙儿认为～＊我没理,大伙儿〔这样〕认为

从背后下手'他们企图～＊从背后下手,他们企图〔那样(做)〕

盖一座楼房'弟弟打算～＊盖一座楼房,弟弟打算〔那样(做)〕

放弃学位'他们迫使我～＊放弃学位,他们迫使我〔这样(做)〕

(7) 不作"V-成"一类复合动词的宾语,例如:

＊觉得成那地方不错|＊认为成我没理|

＊企图成从背后下手|＊打算成盖一座楼房

＊迫使成我放弃学位

3 论元角色的句法功能和范畴约束

上文试图以各别论元角色的语法表现为参考索引(reference index),从语法形式上来限定17种常见的论元角色的疆界。事实上,论元角色的语法表现至少包括句法功能(syntactic function)和范畴特征(categorical feature)两个方面:前者指各个论元角色在句子中各自可以充当什么样的句法成分(比如:主语、宾语、状语),充当什么样的句法形式的主语和宾语(比如:作什么样的介词或动词重叠式的宾语,作什么样的语法形式的主语);后者指各个论元角色通常由什么样的词类范畴来实现,比如:施事、受事通常由名词性成分来实现,致事通常由名词或动词性成分来实现,场所、源点、终点通常由处所性成分来实现。虽然这种论元角色的范畴约束也是识别不同的论元角色的一种重要的语法指标,但是由于不同的论元角色在对词类范畴的选择性方面似乎差别不大、好像都以名词性成分为主要的

① 例句中间左上方的'号表示句法成分易位的边界,那儿通常不允许有较长的停顿。

实现形式；加上我们在这方面的研究工作还做得不够，因此现在暂时付诸阙如。希望今后有机会能够进一步挖掘这方面的语法表现，来弥补这种欠缺。

鸣谢：我的同事詹卫东先生提议和催促我做这个课题，在此谨向他表示诚挚的谢意。

参考文献

陈　平（1994）《试论汉语中三种句子成分与语义成分的配位原则》，《中国语文》第3期。

程　工（1995）《评〈题元原型角色与论元选择〉》，《国外语言学》第3期。

顾　阳（1994）《论元结构理论介绍》，《国外语言学》第1期。

孟　琮等（1987）《动词用法词典》，上海辞书出版社。

汤廷池、张淑敏（1996）《论旨网格、原参语法与机器翻译》，《中国语文》第4期。

徐烈炯（1988）《生成语法理论》，上海外语教育出版社。

徐烈炯（1990）《语义学》，语文出版社。

杨成凯（1986）《Fillmore的格语法理论》，《国外语言学》第1、2、3期。

袁毓林（1998）《汉语动词的配价研究》，江西教育出版社。

朱德熙（1982）《语法讲义》，商务印书馆。

Abraham, W. (ed.) (1978) *Valence, Semantic Case and Grammatical Relation*, John Benjamin's.

Dowty, D. (1991) Thematic Proto-Role and Argument Selection. *Language*, Vol. 67, No. 3.

Fillmore, C. (1968) The Case for Case. *Universals in Linguistic Theory*, (ed.) By Emmon Bach and Robert T. Harms, 1—90. New York: Holt, Rinehart & Winston.《"格"辨》，胡明扬译，《语言学译丛》，第二辑，第1—117页，中国社会科学出版社，1980年。

Fillmore, C. (1977a) The Case for Case Reopened, in p. Cole & J. M. Sadock (eds.) *Syntax and Semantics*, Vol. 8, *Grammatical Relations*, pp. 59—81. Academic Press.

Fillmore, C. (1977b) Topics in Lexical Semantics, in R. W. Cole (ed.) *Current Issues in Linguistic Theory*, pp. 76—138.

Grimshaw, J. (1990) *Argument Structure*. MIT Press.

Gruber, J. (1976) *Lexical Structures in Syntax and Semantics*. North-Holland.

Hale, K. & S. J. Keyser (1991) *On the Syntax of Argument Structure*. MIT Press.

Jackendoff, R. (1990) *Semantic Structure*. The MIT Press.

Steinberg, D. & Jakobovits, L. (1971) *Semantics: An Interdisciplinary Reader in Philosophy, Linguistics and Psychology*. Cambridge University Press.

Williams, E. (1994) *Thematic Structure in Syntax*. MIT Press.

2001年5月初稿,2002年3月改定

(发表于《世界汉语教学》2003年第3期)

汉语谓词的论元结构的描述框架

本文通过十个具体的例子，来说明怎样来建立汉语谓词（动词、形容词、有价名词）的论元结构的描述框架。内容包括(i) 谓词的论元属性（价数），(ii) 这些论元的论旨属性（语义格），(iii) 这些论元角色的词类范畴特性和句法特性（成分功能），(iv) 充当这些论元角色的词语的语义特性，(v) 谓词及其论元的句法配置方式及其典型格式和例句。

0 引 言

袁毓林(2002)指出，动词的论元结构的主要内容应该包括：

(1) 论元属性：确定每一个动词能支配多少个必用论元、多少个可用论元；

(2) 论旨属性：标定这些论元在语义上的功能，即论旨角色；

(3) 语法特征：描写这些论元在句法上的功能和所受到的句法约束；

(4) 语义特征：描写这些论元的动态的语义特征和静态的语义特征；

(5) 配位方式：描写动词及其论元的句法配置方式。

其中，论元属性(argument property)指的是动词所能关联的论元的数目，通俗地说就是价数。为了尽可能精确和细密，我们用袁毓林(1998：100)的"联、项、位、元"这种配价层级作框架。具体地说，联指一个动词在各种句子中所能关联的语义角色不同的名词性成分的数量，项指一个动词在一个句子中所能关联的名词性成分的数量（其中包括通过介词引导的名词性成分），位指一个动词在一个句子中不借助介词所能关联的名词性成分的数量，元指一个动词在一个简单的基础句中所能关联的名词性成分的数量。

论旨属性(thematic property)指的是各个论元的论旨角色,通俗地说就是语义格(case)。

语法特征(grammatical feature)包括句法功能(syntactic function)和范畴特征(categorical feature)两个方面,前者指各个论元在句子中各自可以充当什么样的句法成分(比如:主语、宾语、状语),后者指各个论元通常由什么样的词类范畴来实现(比如:施事、受事通常由名词性成分来实现,致事通常由名词或动词性成分来实现,场所、源点、终点通常由处所词、方位词等处所性成分来实现)。

语义特征(semantic feature)包括动态的语义特征和静态的语义特征两个方面,前者指的是各个论元在述谓结构中表现出来的施动性、受动性等语义特点,后者指实现不同的论旨角色的词语在语义上受到的约束(比如:施事、与事通常由指人名词来实现,受事则既可以由指人名词来实现,也可以由指物名词来实现)。

配位方式(argument selection)指的是依存于同一个动词的各个论元在句子中的共现和选择限制,即怎样构成一个或几个相关的句式,也就是说,动词及其论元角色的句法配置方式。必须注意的是,所谓动词的论元结构实际上指的是动词的某个义项或义位(sememe)的论元结构,也就是说,同一个动词的不同义位可能具有各不相同的论元结构。

我们认为,这些内容对于掌握动词的意义和用法,起着至关重要的作用。甚至可以说,说母语者正是拥有了这些知识,才得以正确地遣词造句和听读理解;一个外族学习者只有了解了这些知识,才能用这些词来造句或理解包含这些词的语句。同样,计算机要理解自然语言,这种结构化的论元结构知识是必不可少的。

本文打算通过十个具体的例子,来说明怎样来建立汉语谓词的论元结构的描述框架。其中,各种论元角色的动态的语义特征,已经在袁毓林(2002)作了总的讨论,不必要在每个动词上重复一下,这里从略。必须说明的是,这里所谓的谓词除了指动词和形容词之外,还包括有价名词。

1. 切：qiē,〈动词〉用刀把物品分成若干部分

〔1〕**论元属性**：二元三位四项六联动词，记作：V：2—3—4—6；或称：二元六系动词，记作 V：2—6，或 V^{2-6}。

〔2〕**论旨属性**：{施事 A，受事 P，结果 R，方式 M，工具 I，处所 L}，记作：V：{A，P，R，M，I，L}；其中，结果 R 是受事 P 和方式 M 的合并，记作：R＝P＋M，如"黄瓜片儿、辣椒丝儿"等。

〔3〕**范畴特性**：均为名词性成分，记作：A，P，R，M，I，L → NP；其中，L 只能是名词和方位词构成的复合处所词，如："案板上，碟子里"。

〔4〕**句法特性**：在基础句中，施事 A 实现为主语；受事 P 和结果 R、方式 M 分别实现为宾语；工具 I 实现为状语中介词"用"的宾语；处所 L 只有在派生句中才能出现，实现为宾语或主语。记作：A → Sub；P，R，M → Obj；I → Prep＋Obj；L → Obj/Sub。

〔5〕**语义特性**：A → NP[＋Human]；M → NP[＋Abstract]，如"片儿、丝儿"等；I→NP[＋Tool]，如"刀儿、菜刀"等；L → NP[＋Place]，如"案板上、碟子里"等。

〔6〕**句法配置**："切"及其论元能构成下列句式：

S1：A＋用 I＋__＋P / R

S2：I＋A＋__＋P / R

S3：P＋A＋__＋M

S4：A＋把 P / R＋__＋L

S5：R＋A＋__＋L

S6：L＋__ 着/了＋R

S7：A＋用 I＋把 R＋__＋在 L

说明：其中，斜撇/表示析取(disjunction)关系。

〔7〕**典型句式**：

S1：他用小刀～黄瓜/黄瓜片

S2：这把刀我～黄瓜/黄瓜片

S3：这根黄瓜你～丝儿/片儿

S4：他把黄瓜片儿～案板上了

S5：黄瓜片儿他～案板上了

S6：案板上～了不少黄瓜片儿

S7：小刚用水果刀把黄瓜片儿全～案板上了

〔8〕**真实例句：**（略）

2. 包：bāo，〈动词〉通过包裹的方式制作食品

〔1〕**论元属性：**二元三位（三项三联）动词，记作：V：2—3(—3—3)；或称：二元三系动词，记作：V：2—3，或 V^{2-3}。

说明：当联的数目等于项时，可以只标记到项；同样，当项的数目等于位时，可以只标记到位；依此类推，当位的数目等于元时，可以只标记到元。也就是说，上面圆括号中的数目是可以省略的。

〔2〕**论旨属性：**{施事 A，结果 R，材料 Ma}，记作：V：{A，R，Ma}。

〔3〕**范畴特性：**均为名词性成分，记作：A，R，Ma → NP。

〔4〕**句法特性：**在基础句中，施事 A 实现为主语，结果 R 实现为宾语，材料 Ma 实现为状语中介词"用"的宾语；在派生句中，结果 R 和材料 Ma 可以实现为大主语(话题)。记作：A，R，Ma → Sub；P，R → Obj；Ma → Prep＋Obj。

〔5〕**语义特性：**A → NP[＋Human]，R → NP[＋Food]，如"饺子、馄饨、粽子"；Ma → NP[＋Grain]，如"面粉、糯米（粉）"等，有时，像"竹叶、竹箬"等不具有[＋Grain]特征的词语也可以充当这种材料格，如："妈妈用竹叶包粽子|这些竹箬我们包粽子吧"。为了区别，前者可以称为内容性(content)材料，记作 Ma(Con)；后者可以称为工具性(tool)材料，记作 Ma(Tol)。

〔6〕**句法配置：**"包"及其论元能构成下列句式：

S1：A＋用 Ma＋__＋R

S2：Ma＋A＋__＋R

S3：R＋A＋__了…

S4：A＋把 Ma＋__了＋R

说明：S4 中的材料 Ma 一般只能是内容性材料 Ma(Con)，而不能是工具性材料 Ma(Tol)。

〔7〕**典型句式**：

S1：妈妈用那些面粉～了几十只饺子
S2：这些馅儿我～馄饨|这些竹箬我们～粽子
S3：饺子我～了一些
S4：他把馅儿～了饺子了

〔8〕**真实例句**：（略）

3. 包：bāo,〈动词〉用纸、布等东西把东西包裹起来

〔1〕**论元属性**：二元三位（三项）四联动词，记作：V：2—3（—3）—4；或称：二元四系动词，记作：V：2—4，或 V^{2-4}。

〔2〕**论旨属性**：{施事 A，受事 P，材料 Ma，方式 M}，记作：V：{A，R，Ma，M}。

说明：这里的材料 Ma 一般只能是工具性材料 Ma（Tol），而不能是内容性材料 Ma（Con）。

〔3〕**范畴特性**：均为名词性成分，记作：A，R，Ma，M → NP。

〔4〕**句法特性**：在基础句中，施事 A 实现为主语，受事 P 实现为宾语，材料 Ma 实现为状语中介词"用"的宾语；在派生句中，受事 P 和材料 Ma 可以实现为大主语（话题），方式 M 和材料 Ma 可以实现为宾语。记作：A，P，Ma → Sub；P，Ma，M → Obj；Ma → Prep+Obj。

〔5〕**语义特性**：A → NP[＋Human]，Ma → NP[＋Thin，Two-dimensions]，如"牛皮纸、塑料布、荷叶、彩纸"等，M → NP[＋Abstract]，如"小包、三角包、双十字"等。

〔6〕**句法配置**："包"及其论元能构成下列句式：

S1：A＋用 Ma＋__＋P
S2：Ma＋A＋__＋P
S3：P＋A＋用 Ma＋__
S4：P＋A＋__＋Ma
S5：P＋A＋__＋M

〔7〕**典型句式：**

S1：他正用牛皮纸～书呢
S2：这张牛皮纸我～词典
S3：这本书你用牛皮纸～
S4：这本书你～牛皮纸
S5：这捆书你～双十字|这些糖你～小包

〔8〕**真实例句：**（略）

4. 调查： diàochá,〈动词〉为了解情况进行考察

〔1〕**论元属性：** 二元三位（三项三联）动词，记作：V：2—3（—3—3）；或称：二元三系动词，记作：V：2—3，或 V^{2-3}。

〔2〕**论旨属性：**{施事 A，受事 P，与事 D，命题 Pn}，记作：V：{A，P，Pn}。

〔3〕**范畴特性：** 施事 A，受事 P，与事 R 为名词性成分，记作：A，P，D → NP；命题 Pn 为谓词性成分或小句，记作 Pn→ VP，S'。

说明：S'代表小句。

〔4〕**句法特性：** 在基础句中，施事 A 实现为主语，受事 P 实现为宾语；与事 D 实现为状语中介词"向"的宾语；在派生句中，与事 D 可以实现为宾语，受事 P 可以实现为状语中的介词"为了"的宾语；命题 Pn 只能实现为宾语。记作：A → Sub；P，D，Pn → Obj；P，D → Prep＋Obj。另外，"调查"是名动词，可以作"作、进行"等形式动词的宾语；记作：V → 进行＋Obj。

〔5〕**语义特性：** A，D → NP[＋Human]；P → NP[＋Abstract]，如"情况、事故原因"等；Pn→ VP，S'[＋Que]。

说明：S'[＋Que]表示小句中包括"有没有、是不是"等疑问形式。

〔6〕**句法配置：**"调查"及其论元能构成下列句式：

S1：A＋向 D＋__＋P / Pn
S2：A＋为了 P＋__ D
S3：为了 P＋A＋__＋D
S4：P＋A＋__＋D

S5：A＋为了 P＋向 D＋进行＋__

S6：为了 P＋A＋向 D＋进行＋__

〔7〕**典型句式：**

S1：警察向司机～事故原因/这辆小公共有没有超载

S2：警察为了这起事故～了许多目击者

S3：为了这起事故警察～了许多目击者

S4：这起事故我们～了十几个幸存者

S5：民警为了这起事故向一些目击者进行～

S6：为了这起事故民警向一些目击者进行～

〔8〕**真实例句：**（略）

5．帮忙：bāngmáng，〈动词〉帮助别人做事

〔1〕**论元属性：**一元二位三项（三联）动词，记作：V：1—2—3（—3）；或称：一元三系动词，记作：V：1—3，或 V^{1-3}。

〔2〕**论旨属性：**｛施事 A，与事 D，范围 Ra｝，记作：V：｛A，D，Ra｝。

〔3〕**范畴特性：**施事 A 和与事 D 为名词性成分，记作：A，D → NP；范围 Ra 为谓词性成分或体词性成分，记作 Ra → VP/NP。

〔4〕**句法特性：**在基础句中，施事 A 实现为主语，与事 D 实现为状语中介词"给"的宾语。由于"帮忙"是离合词，与事 D 还可以插在"帮忙"的中间、作"帮"的间接宾语，"忙"可受"一些、不少"等数量词修饰；这时，范围 Ra 可以作主语，如"我搬家/这件事，小王帮了（我）不少忙"。记作：A，Ra → Sub；P，D → Prep＋Obj。

〔5〕**语义特性：**A，D → NP[＋Human]；Ra → NP/VP[＋Thing/Event]，如"这件事、搬家、盖房子"等。

〔6〕**句法配置：**"帮忙"及其论元能构成下列句式：

S1：A＋给 D＋__

S2：A＋帮了/过＋D＋一些/不少忙

S3：Ra＋A＋帮了/过＋D＋一些/不少忙

〔7〕**典型句式：**

S1：大伙儿给老张～

S2：小王帮了/过老张一些/不少忙

S3：这件事/买房子我帮了/过老张一些/不少忙

〔8〕**真实例句：**（略）

6. 帮忙：bāngmáng，〈形容词〉乐于帮助别人

〔1〕**论元属性：**一元二位（二项二联）形容词，记作：A：1—2（—2—2）；或称：一元二系形容词，记作：A：1—2，或 A^{1-2}。

〔2〕**论旨属性：**{施事A，与事D，范围Ra}，记作：V：{A，D，Ra}。

〔3〕**范畴特性：**施事A，与事D为名词性成分，记作：A，D → NP；范围Ra为谓词性成分或体词性成分，记作 Ra → VP/NP。

〔4〕**句法特性：**在基础句中，施事A实现为主语；与事D实现为状语中介词"对"的宾语；范围Ra可以作大主语（话题）。如"我搬家/这件事，小王对我很帮忙"；也可以嵌在状语"在……上"中，如："在提干这件事上，王部长很帮忙"。记作：A，Ra → Sub；D → Prep＋Obj。

另外，作为形容词的"帮忙"作谓语时经常出现在"很、非常"等程度副词之后。

〔5〕**语义特性：**A，D → NP[＋Human]；Ra → NP[＋Thing/Event]，如"这件事、搬家、盖房子"等。

〔6〕**句法配置：**"帮忙"及其论元能构成下列句式：

S1：A＋对D＋很＿＿

S2：Ra＋A＋对D＋很＿＿

S3：在Ra上＋A＋对D＋很＿＿

〔7〕**典型句式：**

S1：小王对我很～

S2：我搬家/这件事，小王对我很～

S3：在提干这件事上，王部长对我很～

〔8〕**真实例句：**（略）

7. 飞：fēi,〈动词〉飞跑,即通过飞的方式逃跑

〔1〕**论元属性：**一元二位三项（三联）动词,记作：V：1—2—3（—3）;或称：一元三系动词,记作：V：1—3,或 V^{1-3}。

〔2〕**论旨属性：**{施事 A，来源处所 L(S)，目标处所 L(G)},记作：V：{A，L(S)，L(G)}。

〔3〕**范畴特性：**均为名词性成分,记作：A，L(S)，L(G) → NP。

〔4〕**句法特性：**在基础句中,施事 A 实现为主语,来源处所 L(S) 和目标处所 L(G)分别实现为状语中介词“从、到”的宾语;在派生句中来源处所 L(S)可以实现为主语,目标处所 L(G) 可以实现为宾语;当来源处所 L(S)实现为主语时,施事 A 实现为宾语。记作：A，L(S) → Sub;L(G)，A → Obj;L(S)，L(G) → Prep＋Obj。

〔5〕**语义特性：**A → NP[＋Animal，Winged],L(S)，L(G) → NP[＋Place]。

〔6〕**句法配置：**“飞”及其论元能构成下列句式：

S1：A＋_
S2：A＋从 L(S)＋_ 了
S3：(从)L(S)＋_了＋A
S4：A＋_＋L(G) 了
S5：A＋从 L(S)＋_＋L(G)了

〔7〕**典型句式：**

S1：鸽子～了
S2：鸽子从鸟笼里～了
S3：(从)鸟笼里～了一只鸽子
S4：鸽子～(到)屋外了
S5：鸽子从鸟笼里～(到)屋外了

〔8〕**真实例句：**（略）

8. 飞：fēi，〈动词〉由飞跑造成丢失

〔1〕**论元属性**：二元（二位二项二联）动词，记作：V：2（—2—2—2）；或称：二元二系动词，记作：V：2—2，或 V^{2-2}。

〔2〕**论旨属性**：{遭受性主事 Th(P)，施动性受事 P(A)}，记作：V：{Th(P)，P(A)}。

说明："遭受性主事"和"施动性受事"这两个名称听上去很怪，其实正反映了论旨角色的复杂性。像主事 theme，有人干脆就翻译成"客体"；而"施动性受事"则有点儿像是作格(ergative case)。

〔3〕**范畴特性**：均为名词性成分，记作：Th(P)，P(A) → NP。

〔4〕**句法特性**：遭受性主事 Th(P)实现为主语；施动性受事 P(A)实现为宾语，并且强制性地需要有数量词修饰。记作：Th(P) → Sub；P(A) → Obj。

〔5〕**语义特性**：Th(P) → NP[+Human]，P(A) → NP[+Animal，Winged]。

〔6〕**句法配置**："飞"及其论元能构成下列句式：

S1：Th(P)+_+P(A)

〔7〕**典型句式**：

S1：老王～了一只鸽子

〔8〕**真实例句**：（略）

9. 吃：chī，〈动词〉吃别人的东西，即从别人那儿吃东西

〔1〕**论元属性**：三元（三位三项三联）动词，记作：V：3（—3—3—3）；或称：三元三系动词，记作：V：3—3，或 V^{3-3}。

〔2〕**论旨属性**：{施事 A，受事 P，与事 D}，记作：V：{A，P，D}。

〔3〕**范畴特性**：均为名词性成分，记作：A，P，D → NP。

〔4〕**句法特性**：在基础句中，施事 A 实现为主语，受事 P 实现为直接宾语，与事 D 实现为间接宾语。一般来说，受事名词前必须有数量词作修饰语。记作：A → Sub；P → Obj2；D → Obj1。值得注意的是，与事 D 可以通过附加方位词"那儿"而转化为来源处所 L

(G)。例如:“我吃了小王一个苹果→我从小王那儿吃了一个苹果”。

〔5〕**语义特性**: A, D → NP[+Human], P → NP[+Food/fruit],如“馒头、苹果”等。

〔6〕**句法配置**:“吃”及其论元能构成下列句式:

S1: A+__+D+P

S2: A+从 D 那儿+__+P

〔7〕**典型句式**:

S1: 我~了小张一个苹果

S2: 我从小张那儿~了一个苹果

〔8〕**真实例句**:(略)

10. 专政: zhuānzhèng,〈名词〉统治阶级对敌对势力实行的强力统治

〔1〕**论元属性**:二元名词,记作: N: 2,或 N^2。

〔2〕**论旨属性**: {降级施事 dgA, 降级受事 dgP},记作: V: {dgA, dgP}。

说明: dg 是降级述谓结构(down graded predication)的缩写。

〔3〕**范畴特性**:降级施事 dgA,降级受事 dgP 为名词性成分,记作: dgA, dgP → NP。

〔4〕**句法特性**:在基础句中,二元名词“专政”作形式动词“作、进行、实行”等的宾语。“专政”的降级施事 dgA 跟“实行”等形式动词的施事共价,记作 A(=dgA),实现为主语;“专政”的降级受事 dgP 实现为状语中介词“对”的宾语,同时作“实行专政”这个动词性结构的与事,记作 D(=dgP)。上面的叙述,可以记作: A(=dgA) → Sub; N^2 → Obj; D(=dgP) → Prep+Obj。另外,把这种基础句“A(=dgA)+对 D(=dgP)+实行+专政”中的形式动词“实行”换成名词化标记“的”,就变成名词化的偏正词组“A(=dgA)+对 D(=dgP)+的+专政”。

〔5〕**语义特性**: A, P → NP[+Human]。

〔6〕**句法配置:**“专政”及其论元能构成下列句式:

S1:A(=dgA)+对 D(=dgP)+实行+__
S2:A(=dgA)+对 D(=dgP) +的+__

〔7〕**典型句式:**

S1:无产阶级对资产阶级实行~
S2:无产阶级对资产阶级的~

〔8〕**真实例句:**(略)

参考文献

顾　阳 (1994)《论元结构理论介绍》,《国外语言学》第1期。
汤廷池、张淑敏 (1996)《论旨网格、原参语法与机器翻译》,《中国语文》第4期。
袁毓林 (1998)《汉语动词的配价研究》,江西教育出版社。
袁毓林 (2002)《论元角色的层级关系和语义特征》,《世界汉语教学》第3期。
袁毓林 (2003)《一套动词的论元角色的语法指标》,《世界汉语教学》第3期。

2001年5月初稿,2004年6月改定

论元结构和句式结构
互动的动因、机制和条件

——表达精细化对动词配价和句式构造的影响

本文首先讨论动词配价学说、论元结构理论和句式语法的有关理论和观念，比较它们对于动词和句式的关系的不同认识以及相应的处理方法。接着，指出表达的精细化等语用动机促动了句式套用和词项代入，这又引发了动词和句式的互动，其结果是动词改变其论元结构来适应句式意义和句式构造的需要。然后，指出句式套用和词项代入的认知基础是隐喻投射和完形包装，强调以归纳动词的句法组配模式为逻辑起点，可以超越动词配价和句式构造之间的循环论证。特别强调与事插入、施事删除等规则是动词和句式互动的具体机制，在一定的句法、语义条件下启动这些规则就可以使动词衍生出符合句式要求的论元结构。最后，说明动词和句式的对应关系是有理据的、但又是不可预测的，动词和句式互动背后的逻辑机制是追求动因解释的归因推理。

1 如何解释论元增容：
从动词配价走向句式配价

1.1 动词配价学说和论元结构理论

朱德熙(1978)关于汉语动词“向”的研究，直接开启了上个世纪八十年代至今二十多年的汉语动词配价(valence)研究的热潮。期间，我们的动词配价研究不仅接受了德国和法国配价语法和依存语法的有关理论和方法，而且吸收了在生成语法背景上提出的格语法、论元结构理论等的有关观念和分析技术。① 动词配价研究的主要目的是：通过刻画动词和相关的名词性成分之间的支配关系及其句法

---

① 详见袁毓林(1998)，第48—98页。

配列(syntactic arrangement),来解释句法结构的合格性、并说明句法结构跟语义结构之间的映射关系。[①] 例如:

(1) a. 国家主席江泽民会见了美国总统克林顿

b. *国家主席江泽民会面了美国总统克林顿

(2) a. 国家主席江泽民跟美国总统克林顿在雅加达会面

b. *国家主席江泽民跟美国总统克林顿在雅加达会见

由于二价动词"会见"可以支配施事和受事两个论元,并且受事论元只能实现为宾语,因而(2b)是不合格的表达;由于准二价动词"会面"可以支配施事和与事两个论元,并且与事论元只能实现为介词宾语、不能实现为宾语,因而(1b)是不合格的表达。

可见,汉语动词配价研究奉行的是动词中心论,其核心的思想有两点:

(i) 动词决定多少种和什么样的从属成分(或称补足语)跟它共现;

(ii) 动词具有n元关系,等待着一定数目和类型的论元来填充。

这论元数目就是价数,这论元的类型主要指论旨角色(即语义角色,或语义格,俗称价类)。

这种观念正好顺应了美国语法学研究中的词汇主义(lexicalism)思潮。采取词汇主义这种研究路子的学者相信:[②]

(i) 动词的意义跟句法框架相关,动词的句法范畴框架(N+V+N…)可以从动词的词汇语义上预测。也就是说,句法是词项要求的实现(投射),句法框架是动词意义的表层反映。比如,Jackendoff 等把这种思想提炼为动词组合的透明规则:动词的意义就是一个谓词带着一组固定的论元,并造成一个命题。

① 详见袁毓林(1998),第87、96页。

② 详见 Goldberg (1995),p. 7—19;Levin & Rappaport (1997),p. 487—489。

(ii) 从语义角色或论旨阵列上预测显性的句法。比如，Levin(1985)认为：普遍的连接规则(linking rule)把语义的论元结构映射为显性的补足语结构。

这种思想的集中体现就是所谓的论元结构理论，其中，Jackendoff (1972：43)提出了著名的论旨阶层(thematic hierarchy)，意思是不同的论旨角色是按照阶层的形式排列的。可以表示为(当然，不同的学者对论旨阶层上不同论旨的先后次序有不同的认识)：

施事 > 处所/终点/起点 > 客体

Larson (1988：382)提出了著名的论旨指派原则：

如果一个动词 α 决定若干个论旨角色 θ_1，θ_2，…，θ_n，那么将论旨阶层上最低的那个论旨角色指派给句子成分结构(constituent structure)中位置最低的那个论元；然后，依次类推地指派其余的论旨角色。

这样，论旨阶层跟深层结构上的成分结构就是一种直接的映射(direct mapping)；凡是表层结构中论元位置跟预先设计好的论旨关系次序不对应，就必须用句法上的移位来处理。[①] 比如，在词库中给出动词 put 的论旨关系和次范畴属性两种描述，那么就很容易推导出由 put 构成的句子：[②]

(3) a. *put* (Agent (Theme (Location)))

b. *put*, V[NP(Agent)[__, NP(Theme), PP(Location)]]

c. John put the book on the table.

这种词汇主义的研究路线，符合弗雷格(Frege)提出的意义的组合性(compositionality)原理：一个语言中的每一个表达式的意义是其直接构成成分的意义和用以联结这些成分的句法规则的函项(function)。这样，如果把动词的配价性质搞清楚了，那么句子的基

① 中文介绍详见顾阳(1994)，第4—5页。

② 下面的举例是根据顾阳(1996：4)改编的。

本构造和语义解释也就基本抓住了。也正因为如此，汉语动词配价研究曾经并仍然得到中文信息处理专家的青睐。

1.2　动词变价和论元增容过程

动词配价分析在方法论上是属于自底向上式的（bottom-up）还原主义（reductionism）。这种分析方法虽然简捷明快，但是它并不总是奏效的；最突出的一点是：它不能很好地解释从动词的语义和配价上无法预测的句式构造。例如：[①]

(1) a. 老王扔我一包烟。
　　b. 他吃了我一个苹果。
(2) a. 他烂了几个橘子。
　　b. 他坐了一屁股泥巴。
(3) a. 他摸了一手油污。
　　b. 他急了一身汗。
　　c. 这事急了他一身汗。

例(1)中的二价动词“扔、吃”带了施事（agent，简称 A）、受事（patient，简称 P）和与事（dative，简称 D）三个论元，这与事论元本来不是这两个动词的语义所蕴涵的语义角色；例(2)和例(3b)中的一价动词或形容词“烂、坐、急”带了当事（experiencer）和客体（theme）或结果（resultive）两个论元，同样这客体或结果论元本来不是这三个动词（包括形容词）的语义所蕴涵的语义角色。本来二价动词“摸”可以带施事和受事两个论元，但是在例(3a)中它带了施事和结果两个论元，同样这结果论元本来不是动词“摸”的语义所蕴涵的语义角色。在例(3c)中，一价动词“急”居然带了致事（causer）、受事和结果三个论元，显然，受事和结果两个论元本来不是“急”的语义所蕴涵的语义角色。

对此，着眼于动词配价的学者的自然的反应是：仍然把这种句法现象归结为动词本身，常规的做法是把它处理为由于词义变化带

① 例子和说明参考沈家煊(2000)，第 291 页。

来的配价变化。比如，马庆株(1983：107)认为：在“扔我一个球”一类句子中，“扔”类动词本身没有给予意义，经常用作二价动词；只是在双宾语构造里才具有给予意义，成为三价动词。马庆株(1998)进一步指出，价数固定的动词是定价动词，价数不固定的动词是变价动词；价数受义项的影响，如“吃、扔”一般表现为二价，在一定条件下(双宾构造中)又会表现为三价(第 286 页)；配价成分数量的变化是这种变价动词的形式标志(第 284 页)。

这种所谓的动词变价现象，在英语中也是屡见不鲜的。例如：①

(4) a. Sally baked her sister a cake.
 b. Joe painted Sally a picture.
 c. Joe cleared Sam a place on the floor.

(5) a. Pat threw Chris the ball.
 b. Chris kicked Pat the ball.
 c. Pat hit Chris the ball.

(6) a. Dan talked himself blue in the face.
 b. Sam carefully broke the eggs into the bowl.
 c. He sneezed the napkin off the table.

例(4)中的 bake、paint、clear 和例(5)中的 threw、kicked、hit 是二价动词，只能带施事和受事两个论元，这里却多带了一个与事论元。例(6a)中的 talk 是一价动词，只能带施事一个论元，这里却多带了一个受事论元和一个结果论元；例(6b)中的 break 是二价动词，只能带施事和受事两个论元，这里却多带了一个处所论元；例(6c)中的 sneeze 是一价动词，只能带施事一个论元，这里却多带了一个受事论元和一个处所论元。

对此，Larson (1990) 认为，上例中的 bake、hit 等动词经历了一个词汇派生(derivation)过程，这个过程被称作论元增容(argument augmentation)，它可以在一定的条件下给动词的论元结构增加新的论元。比如，英语及物动词的论元结构增加受益者(bene-

① 例(4)—(6)引自 Goldberg (1995)，p. 9，21，22，34，35，141，143。

factive)和目标(goal)论元的词汇规则(lexical rule)可以具体地表示如下：

(7) 增加受益者(可选)：向动词α的论旨网格(θ-grid)中增加受益者论旨角色。

条件：动词α表示制作(creation)或准备(preparation)事件(event)。

结果：客体为受益者提供了利益。

(8) 增加目标(可选)：向动词α的论旨网格中增加目标论旨角色。

条件：动词α表示运动(motion)事件，其中施事向客体发出一个射体轨道。

可见，论元增容是受词汇和语义条件限制的。比如，增加受益者要求动词表示制作或准备意义，动词所支配的客体所表示的必须是成事宾语，这种宾语是通过动词所描述的事件创造出来的，它可以使新增加的受惠者论元得益。根据词汇规则(7)和(8)，可以把动词 bake、hit 的论元增容过程表示如下：[①]

(9) a. bake：{θ施事者，θ客体}

↓

论元增容(增加域内论元：受惠者)

↓

b. bake：{θ施事者，θ客体，θ受惠者}

(10) a. hit：{θ施事者，θ客体}

↓

论元增容(增加域内论元：目标)

↓

b. hit：{θ施事者，θ客体，θ目标}

(9a)(10a)是 bake、hit 固有的论元结构，(9b)(10b)是论元增容后

① 详见 Larson (1990)，p. 615—618；顾阳(1999)作了很好的介绍，并作了一定的引申和发挥，第 81—82 页。

bake、hit 的论元结构。(10b)新增加的受惠者论元可以用介词 for 引导,从而投射成与格结构(如: Mary baked a cake for John.)。因为介词 for 本身含有受惠义,跟新增加的受惠者论元意义相重;所以这种与格结构可以经过被动化处理,从而得到双宾语结构(如: Mary baked John a cake.)。

这种做法的实质就是碰到新的用法就给动词增加意义,但是,调用(7)(8)这种可选性的词汇规则的动因和条件并不明确。如果这种变价动词为数不多,那么或许可以把这种变价用法归结为是这少数动词的词汇特异性(lexical idiosyncrasy)。可事实是,这种变价用法是比较普遍的,不仅二价动词在特定句式中可以带三个论元,而且一价动词在特定句式中也可以带三个论元。这就需要一种更有概括性和解释力的理论模型来处理这类现象。

1.3 句式语法和句式配价

在 Fillmore、Kay、O'Connor、Lakoff、Brugman、Lambrecht、Langacker 等学者关于句式(constructions)的工作的影响下,Goldberg (1995)提出了系统的句式语法(construction Grammar)的思想和分析方法。这种句式路线(constructional approach)在本质上是反对词汇路线(lexical approach)的,其中心观点是:英语的基础句(basic sentences)是句式的实例(instances),句式是一种"形式+意义"配对,它独立存在于特定的动词。即句式自己负载意义,独立于句中词项的意义;也就是说,句子的语义结构及其形式表达是由独立于其构成词项的句式造成的(p. 1)。这跟 Chomsky (1981, 1992)等认为句法构造(syntactic constructions)是由普遍原则的互动作用而造成的附带现象(epiphenomenal)的观点迥然不同。这样,上文讨论的动词变价和论元增容就不必归结为同一个动词有几种不同的意义(sense),而是可以非常节俭地把同一动词在不同句式中的意义差别归结为特定的句式。

Goldberg (1995)对句式下的定义是:如果一个"形式—意义"配对(form-meaning correspondences)的形式或意义方面的特性不能从其构成成分或其他句式上推导出来,那么它就是一个句式(p. 4)。

并且认为，简单的小句结构跟反映人类基本经验的语义结构直接相关，句式所涉及的基本的论元结构是跟动态的场景（有经验基础的格式塔）相关的（p. 5）。从而构建了一种解释性的、而不是生成性的单层次的（monostratal）语法理论。在怎样看待动词和句子的论元结构关系上，这种句式路线跟词汇路线最大的不同点是，它强调动词跟句式相关但各自独立，框式结构（skeletal constructions）可以提供论元，比如双宾语结构（double object constructions）可以允准与事论元。于是，二价的 bake、cook 等制作（create）动词可以进入双宾语结构。这样，句子中论元成分之间的 n 元关系直接跟框式结构相联系，动词只跟少量的基础义项相联系，这些意义一定能整合进句式意义中（p. 11）。当一个动词出现在不同的句式中时，整个句式的意义及限制是不同的。这种不同不必归结为动词的不同义项，可以更节俭地归结为这些不同的句式本身（p. 13）。由于句法框架直接跟意义相联系，并且独立于出现于其中的动词（p. 19）；因而关于语义的组合性原理可以表达成如下这种弱形式：一个表达式的意义是构成词项的意义和句式意义的整合（p. 16）。

在这种句式语法思想的影响下，沈家煊（2000）毅然地把配价看作是句式的属性；并指出：句式配价指抽象的句式配备的、与谓语动词同现的名词性成分的数目和类属（指施事、受事、与事、工具等）。这样，“他扔我一个球”属于三价句式，跟“我送他一本书”一样有施事、受事和与事三个论元，尽管“扔”的词义只涉及两个参与角色（participant role）；“（她结婚）你送什么？”属于二价句式，包含施事和受事两个论元，尽管“送”的词义涉及施事、受事和与事三个参与角色。同样，“王冕死了父亲”属于二价句式，跟“他丢了一枚戒指”一样包含两个论元，尽管“死”的词义只涉及一个参与角色（第 293—4 页）。用这种思想来解释 §1.2 中的例（1）—（6）这类论元增容的句子，倒不失为一种简捷的办法。问题是，这种句式的配价能力是由什么决定的呢？沈先生的回答是，句式的配价或论元主要是由句式的整体意义所决定的，“王冕死了父亲”所属的句式的整体意义要求这个句式有两个论元，“王冕的父亲死了”所属的句式的整体意义只要求这个句式有一个论元（第 294 页）。我们认为，问题没有这么简单

和轻松。因为，接下来的问题该是：(1) 句式的整体意义是由什么决定的？(2) 句式对进入其中的动词的选择限制条件是什么？如果不能很好地解决这两个问题，那么句式语法和句式配价路线就不会比词汇语法和动词配价路线高明多少。充其量也只是把动词变价和论元增容的球踢到了句式这个楼上(kick upstairs)。

2　表达精细化和句式套用、词项代入

2.1　句式意义从何而来？

沈家煊(2000)强调，句式的配价或论元主要是由句式的整体意义所决定的(第 294 页)。这也许是不错的，比如，表示转让(transfer)意义的句式要求施事、受事和与事三个论元，而不管进入其中的动词是二价的还是三价的；表示丧失(lose)意义的句式要求当事(experiencer，简称 E)和客体(theme，简称 Th)两个论元，而不管进入其中的动词是二价的还是三价的。例如：

(1) NP (A)＋V＋NP (D)＋NP (P)

a. 老张 送 小王 一本词典

b. 老刘 卖 小孙 一支钢笔

c. 小平 搛 奶奶 一块鱼排

d. 小明 扔 小华 一个好球

e. 老张 抢 小王 一本词典

f. 老刘 买 小孙 一支钢笔

g. 小平 吃 奶奶 一块鱼排

h. 小明 用 小华 一张宣纸

(2) NP (E)＋V＋NP (Th)

a. 王冕〔七岁时〕失去了 父亲

b. 王冕〔七岁上〕死了 父亲

c. 王大爷 丢了 一串钥匙

d. 王大爷 掉了 一串钥匙

e. 王大爷 丢了 一只鸽子

f. 王大爷 飞了 一只鸽子

g. 我家 损失了 一筐苹果

h. 我家 烂了 一筐苹果

i. 我家 报废了 一台电视

j. 我家 被偷了 一台电视

例(1)中的"送、卖、抢、买"是三价动词,而"搛、扔、吃、用"是二价动词;但是,三价句式"NP (A)＋V＋NP (D)＋NP (P)"使得它们都能跟三个论元发生句法、语义关系。至于为什么这种句式是三价的,显然不能归结为其中的谓语动词(因为,其中既有三价动词、也有二价动词),而是要归结到这种句式所具有的转让意义——转让关系要涉及到转让物(即受事)、让出者(即施事)和接受者(即与事)。例(2)中的"失去、丢、损失"是二价动词,而"死、掉、飞、烂、报废"是一价动词;但是,二价句式"NP (E)＋V＋NP (Th)"使得它们都能跟两个论元发生句法、语义关系。至于为什么这种句式是二价的,显然不能归结为其中的谓语动词(因为,其中既有二价动词、也有一价动词),而是要归结到这种句式所具有的丧失意义——丧失关系要涉及到丧失物(即客体)和受害者(即当事)。

现在的问题是,句式"NP (A)＋V＋NP (D)＋NP (P)"的转让意义是从哪儿来的,句式"NP (E)＋V＋NP (Th)"的丧失意义是从哪儿来的? 显然,词类(形式类)序列"NP＋V＋NP＋NP"和"NP＋V＋NP"本身是不可能产生出转让和丧失之类的句式意义的。一种最有可能的答案是:这种能决定句式配价的句式意义是由动词的论元结构提供的,动词的论元结构中各论元角色之间的语义关系的抽象化为有关句式提供了最初的意义。例如:

(3) a. 送:{送者,送物,受者}

b. 卖:{卖者,卖物,买方}

c. 抢:{抢者,抢物,被抢者}

d. 买:{买者,买物,卖方}

e. V:{施事,受事,与事}

这四个动词的词汇意义都涉及三个参与角色(participant role),如果

对这些参与角色进行概括，那么送者、卖者、抢者、买者等都包含施动性(causation)，因而可以抽象为施事；送物、卖物、抢物、卖者都包含受动性(causally affected)，因而可以抽象为受事；受者、买方、被抢者、卖方等都包含参与性(participant in)，因而可以抽象为与事。"送、卖"等表达的是受事从施事方转移到与事方，可以概括为给予；因此，当它们跟受其支配的论元实现为"NP＋V＋NP＋NP"之类的句法形式时，这种句式自然地具有给予这种句式意义。"抢、买"等表达的是受事从与事方转移到施事方，可以概括为取得；因此，当它们跟受其支配的论元实现为"NP＋V＋NP＋NP"之类的句法形式时，这种句式自然地具有"取得"这种句式意义。给予和取得都涉及受事在施事方和与事方之间转移，只是方向相反；因此，可以进一步概括为转让。于是，句式"NP (A)＋V＋NP (D)＋NP (P)"自然地从其核心动词的论元结构上获得了"转让"这种句式意义。

2.2 句式套用和词项代入

一方面，由于句式意义是由动词的论元结构带来的，因而表示不同意义的句式对进入其中的动词在语义上有严格的选择限制。比如，"NP (A)＋V＋NP (D)＋NP (P)"句式要求其中的动词必须是表示给予或取得等转移意义的，"NP (E)＋V＋NP (Th)"句式要求其中的动词必须是表示丧失意义的。但是，另一方面，典型动词的论元结构被结构(或句法型式，syntactic configuration)包装之后，这个结构(或称句式)也就获得了原型的格式意义；并且，句式作为一种形式和意义的配对，具有相当的模塑性，它能把那些在语义上跟句式意义不同、但是又不相抵触的动词吸收进来。例如：

(1) NP (A)＋V＋NP (D)＋NP (P)

a. 大张 扔 小刘 一包香烟 ← a′. 大张 给 小刘 一包香烟

b. 小平 灌 李伟 一杯白酒 ← b′. 小平 给 李伟 一杯白酒

c. 小明 踢 小华 一个斜线球 ← c′. 小明 给 小华 一个

斜线球

d. 玉芳 孝敬 公公 一条香烟 ← d′. 玉芳 给 公公 一条香烟

e. 李铎 吃 了小邵 一个苹果 ← e′. 李铎 拿 了 小邵 一个苹果

f. 玲玲 只穿过 姥姥 一件毛衣 ← f′. 玲玲 只拿过 姥姥 一件毛衣

g. 老刘 抽 了 小孙 一支香烟 ← h′. 老刘 拿 了 小孙 一支香烟

h. 小芳 花 了 奶奶 一百块钱 ← g′. 小芳 拿 了 奶奶 一百块钱

i. 小明 糟蹋 了 我 好几张宣纸 ← i′. 小明 拿 了 我 好几张宣纸

j. 王平 坑 了 爸爸 一千块钱 ← j′. 王平 拿 了 爸爸 一千块钱

(2) NP (E)＋V＋NP (Th)

a. 王冕 死 了 父亲 ← a′. 王冕 失去 了 父亲

b. 王大爷 飞 了 一只鸽子 ← b′. 王大爷 失去 了 一只鸽子

c. 老王 烂了 几个橘子 ← c′. 老王 失去 了 几个橘子

d. 我家 报废 了 一台电视 ← d′. 我家 失去 了 一台电视

e. 我家 被偷 了 一台电视 ← e′. 我家 失去 了 一台电视

在例(1)中,“扔”指把东西用扔的方式给别人、“灌”指把液体倒进人嘴里,它们具有比较明显的给予性转移意义;所以,可以套用双宾语句式“NP (A)＋V＋NP (D)＋NP (P)”,来表示施事主动地使受事转移到与事方。跟由“送、给”等典型的给予义动词构成的双宾语句在句式意义上的差别是:这种双宾句并不表示受事原来在施事方,而后者则包含受事原来在施事方这种意义。“踢”本来指抬起腿用脚

撞击，在用踢的方法传球的场景知识(scenes knowledge)的影响下，也临时含有给予意义；"孝敬"指把物品献给长者以示敬意，本来就包含一定的给予意义。因此，它们可以套用双宾语句式来表示给予性转移意义。"吃(苹果)"、"穿(毛衣)"本来是消费行为，但是当消费的是别人的东西时，也就等于是从别人那儿(与事方)得到了这种消费品。因此，可以套用双宾语句式"NP (A)＋V＋NP (D)＋NP (P)"，来表示施事从与事方取得某种消费品。跟由"抢、买"等典型的取得义动词构成的双宾语句在句式意义上的差别是：这种双宾句并不表示施事的取得行为一定是主动的(即可以是主动的，如"吃"类双宾语句；也可以是无所谓主动或被动的，如"穿"类双宾语句)，而后者则表示施事一定是主动地实施取得这种行为。"花(钱)"、"抽(烟)"、"糟蹋"本来指耗费或损坏财物并蕴涵失去意义，但是当施事者耗费或损坏别人的财物时，在某种意义上说是从别人那儿(与事方)得到了这种财物(受事)；"坑"指用狡猾、狠毒的手段使人受到损害，这在某种意义上讲也是从别人那儿(与事方)得到了利益(哪怕只是精神上的)。因此，也可以套用双宾语句式"NP (A)＋V＋NP (D)＋NP (P)"，来表示施事主动地使与事方失去财物或利益、并使这种财物或利益转移到与事方。在例(2)中，"(亲人)死(亡)、(宠物)飞(走)、(水果)(腐)烂、(电器)报废"，这对于个人和家庭来说都是一种损失；因此，可以套用表示失去意义的句式"NP (E)＋V＋NP (Th)"，来表示当事失去了客体并由此而造成了损失。有意思的是，我们在《儒林外史》(上海古籍出版社，2000 年)中，找到了类似(2a—a')这种平行的实例：

(3) 这人姓王名冕，……七岁上死了父亲，……。(第 1 回)

(4) 这虞博士三岁上丧了母亲，太翁在人家教书，……。(第 36 回)

"死"是一价的不及物动词，套用了二价的及物动词"丧"的用法。特别要指出的是，在古代汉语中，"丧"有及物和不及物两种用法。例如：(引自《古汉语常用字字典》，第 246 页，商务印书馆，1993)

(5)〔徐〕偃王行仁义而丧其国。(韩非子·五蠹)

(6) 寻程氏妹丧于武昌。(陶潜《归去来兮辞序》,寻：不久)

"丧"作及物动词用时,表示"失去"意义,如(5)所示;作不及物动词用时,表示"死亡"意义,如(6)所示。绝妙的是,在例(4)中,这两种意义好像是兼而有之。

值得注意的是,对于例(1)(2),一方面,我们固然可以说是:"扔、吃"类动词套用了"送、给"类动词惯用的双宾语句式"NP (A)+V+NP (D)+NP (P)","死、飞" 类动词套用了"失去、损失"类动词惯用的"NP (E)+V+NP (Th)"句式,从而凸现(profiling)了这些动词的意义中隐藏着的给予意义。但是,另一方面,我们也可以说是"扔、吃"类动词代换了典型的"送、给"类动词、而进入双宾语句式"NP (A)+V+NP (D)+NP (P)","死、飞" 类动词代换了典型的"失去、损失"类动词、而进入"NP (E)+V+NP (Th)"句式。也就是说,在意义上更为具体的动词代替意义相对抽象的上位动词,具体的下位动词作为抽象的上位动词的一个实例(instance)而进入本来由上位动词主导的句式,从而在表示给予/取得性转移意义的同时,还表示给予的方式扔、灌、踢、孝敬等,或者还表示取得的方式吃、穿、抽、花、糟蹋、坑等;在表示失去意义的同时,还表示失去的方式死亡、飞翔、腐烂、报废、甚至是被盗等。这就是词汇意义和句式意义互动的一个侧面。

2.3 动词代入的语用动因：表达的精细化

根据上面的讨论,句式套用和动词代入是造成动词的配价跟句式配价不一致的一个主要的原因。比如,"扔、灌、踢、孝敬、吃、穿、抽、花、糟蹋、坑"等二价动词可以进入三价的双宾语句式,"死、飞、烂、报废" 等一价动词可以进入二价句式。如果动词的配价跟句式的配价不一致,那么一定会造成动词的参与角色跟句式的论元在数量和类型上的配合不适当,简称角色错配(role mismatches)。例如：

(1) 扔：{扔者,〔受扔者〕,扔物} 如：大张正～手榴弹呢

↓　↓　↓

A＋V＋D ＋ P 如：大张～〔给〕小刘一包香烟

(2) 踢：{踢者，〔接受者〕，踢物} 如：这头牛老～人、孩子们正～足球呢

↓ ↓ ↓

A＋V＋D ＋ P 如：小王～〔给〕我一个斜线球

(3) 吃：{吃者，〔被吃者〕，吃物} 如：我～了一个橘子、她～吃父母

↓ ↓ ↓

A＋V＋D ＋ P 如：我～了小王一个橘子

(4) 坑：{坑者，被坑者，〔被坑物〕} 如：这个鱼贩子老～新顾客

↓ ↓ ↓

A＋V＋D ＋ P 如：这个骗子～了我一大笔保证金

(5) 死：{〔受损者〕，死者} 如：他的父亲～了、张家的小狗～了

↓ ↓

E＋V＋Th 如：他～了父亲、张家～了一只小狗

(6) 飞：{〔受损者〕，飞者} 如：他的鸽子～了、老张的小鸟～了

↓ ↓

E＋V＋Th 如：他～了一只鸽子、老张～了一只小鸟

(7) 烂：{〔受损者〕，烂物} 如：他的苹果～了、张家的白菜全～了

↓ ↓

E＋V＋Th 如：他～了几个苹果、张家～了

一窖白菜

(8) 报废：{〔受损者〕，报废物} 如：他家的电视～了、公司的电脑～了

↓ ↓

E＋V＋Th 如：他家～了一台电视、公司～了一台电脑

从上面的举例可以看出，二价动词“扔”本来只能支配施事（扔者）、受事（扔物）两个论元；但是，进入三价句式“NP (A)＋V＋NP (D)＋NP (P)”后，使潜在的受扔者可以实现为与事论元。二价动词“踢”本来只能支配施事（踢者）、受事（踢物）两个论元，只有用在踢足球等场景中，才可能隐含着接球者这种与事角色；二价动词“吃”本来只能支配施事（吃者）、受事（吃物）两个论元，只有用在从别人那儿吃什么东西的场景中，才可能隐含着被吃者这种与事角色；但是，进入三价句式“NP (A)＋V＋NP (D)＋NP (P)”后，“踢”和“吃”都可以支配施事、与事、受事三个论元。二价动词“坑”本来只能支配施事（坑者＝骗子）、受事（被坑者＝受害者）两个论元，只有用在从别人那儿骗取财物等场景中，才可能隐含着被骗的财物这种受事角色；并且，在这种场景下，原来的受事（被坑者＝受害者）论元转变为与事论元。一价动词“死、烂、报废”本来只能支配一个客体（死者、烂物、报废物）论元，一价动词“飞” 本来只能支配一个施事论元（飞者）；但是，当它们进入二价句式“NP (E)＋V＋NP (Th)”之后，就额外多出一个当事（受害者）论元。

综上所述，句式套用和动词代入造成了角色错配。角色错配的实质是，动词的论元结构跟句式的论元结构的不一致，并且是句式的论元结构压倒(override)了动词的论元结构。那么，为什么动词要迁就句式往火坑里跳呢。这是受表达精细化这种语用动机的强力驱使而促成的。比如，为了具体地表示给予或取得的方式，就用“扔、灌、踢、孝敬、吃、穿、抽、花、糟蹋、坑”等动词代入“送、给”类动词擅场的“NP (A)＋V＋NP (D)＋NP (P)”句式；为了具体地表示失去的方式，就用“死、飞、烂、报废”等动词、甚至是“被偷”一类动词性结构代入“失去”类动词擅场的“NP (E)＋V＋NP (Th)”句式。再如：

(9) a. 床上 躺着 一个病人 ← b. 床上 有 一个病人

(10) a. 楼上 住着 几个留学生 ← b. 楼上 有 几个留学生

(11) a. 园子里 种了 两棵枣树 ← b. 园子里 有 两棵枣树

(12) a. 墙上 挂了 一幅山水画 ← b. 墙上 有 一幅山水画

在例(9)—(12)中,a 式和 b 式都表示存在;但是,b 式表示抽象的存在,a 式通过用具体的动词性结构代换抽象的存在动词“有”之后,指定了具体的存在方式。[①]

这就是说,表达精细化(elaboration)这种语用动机,促动了句式套用和动词代入,最终造成动词和句式在论元结构上的不一致,以至很难用动词的论元结构来解释句子的结构方式及其语义表达。例如:[②]

(13) a. 一个月的工资全被他喝了

← b. 一个月的工资全被他花了

(14) a. 一个月的工资全被他喝了猫儿尿了

← b. 一个月的工资全被他花在喝酒上了

(15) a. 他把一个月的工资全玩了

← b. 他把一个月的工资全花了

(16) a. 他把一个月的工资全玩了麻将了

← b. 他把一个月的工资全花在打麻将上了

(17) a. 你这样做会被别人笑掉大牙的

← b. 你这样做会被别人耻笑的

(18) a. 我可是想死你啦(=了+啊)

← b. 我可是真想你啊

(19) a. 李四被后边的司机按了一喇叭

← b. 李四被后边的司机警告了一下

---

① 参考朱德熙(1981/1990),第 11 页。

② 例(19a)出自 Tan, F. (谭馥) (1991: 166) *Notion of Subject in Chinese*. Ph. D. dissertation, Stanford University, CA. 转引自潘海华(1997)第 6 页。例(20a)引自潘海华(1997)第 6 页。

(20) a. 老师被学生贴了大字报

← b. 老师被学生批判了

用"喝、喝了猫儿尿、玩、玩了麻将"代替"花"、用"笑掉大牙、想死了"代替"耻笑、真想",用"按了一喇叭、贴了大字报"代替"警告了一下、批判了",造成了(13a)—(20a)这种难以用核心动词的句法、语义功能来解释的特殊句式。

3　句式对动词的选择限制条件

3.1　句式的不完全能产性

如果句式具有配价能力,那么它可以自由地指派(assign)论元;于是,特定句式对某种语义类别的动词应该具有相当的开放性。但是,事实上,正如 Goldberg (1995: 120)所指出的,许多句式只是在一定程度上具有能产性(are used somewhat productively),即具有部分的能产性(partial productivity),而不是完全的能产性(full productivity)。例如:①

(1) a. Joe gave $5 to the earthquake relief fund.

→ b. Joe gave the earthquake relief fund $5.

(2) a. Joe donated $5 to the earthquake relief fund.

→ b. *Joe donated the earthquake relief fund $5.

(3) a. Joe told the news to Mary.

→ b. Joe told Mary the news.

(4) a. Joe whispered the news to Mary.

→ b. *Joe whispered Mary the news.

(5) a. Joe baked a cake for Mary.

→ b. Joe baked Mary a cake.

(6) a. Joe iced a cake for Mary.

① 例子和说明,根据 Goldberg (1995), p. 121、130—131 改编。

→ b. * Joe iced Mary a cake.

(7) a. She threw a cannonball to him.

→ b. She threw him a cannonball.

(8) a. She blasted a cannonball to him.

→ b. * She blasted him a cannonball.

(9) Sally permitted / allowed / * let / * enabled Bob a kiss.

(10) Sally refused / denied / * prevented / * disallowed / * forbade Bob a kiss.

从例(1)—(8)可以看出,双及物句式(ditransitive construction)对动词的选择是难以预测的。比如,同样是给予义动词,give 可以,而 donate 不行;同样是言说义动词,tell 可以,而 whisper 不行;同样是制作(creation)义动词,bake 可以,而 ice 不行;同样是弹道运动(ballistic motion)义动词,threw 可以,而 blast 不行。从例(9)和(10)可以看出,同样是许可(permission)义动词,permit, allow 可以,而 let, enable 不行;同样是拒绝(refusal)义动词,refused, deny 可以,而 prevented, disallow, forbid 不行。

汉语的情况也一样,句法、语义性质很接近的一组动词,不一定都能进入相同的句式。例如:

(11) a. 我吃了弟弟一个苹果

b. *我啃了弟弟一个猪手

c. *我嚼了弟弟一根香蕉

d. *我尝了弟弟一口蛋汤

(12) a. 我穿过舅舅一件毛衣

b. 我戴过舅舅一顶帽子

c. *我披过舅舅一件斗篷

d. *我围过舅舅一条纱巾

(13) a. 动物园飞了一只鹦鹉

b. *动物园蹿了一只豹子

c. *动物园蹦了一只袋鼠

d. *动物园跳了一只猴子

e. ＊动物园溜了一只狐狸
f. ＊动物园走了一只孔雀
g. ＊动物园滚了一只猪獾
h. ＊动物园爬了一只乌龟
i. ＊动物园游了一只白鹅

同样是二价的摄食动词，"吃"可以进入三价句式，但"啃、嚼、尝"不能；同样是二价的服饰动词，"穿、戴"可以进入三价句式，但"披、围"不能；同样是一价的移动动词，"飞"可以进入二价句式，但"蹿、蹦、跳、溜、走、滚、爬、游"不能。

对于这种句式的不完全能产现象，如果不能找到合理和充分的解释；那么，句式作为一种独立自主的语法实体（跟词汇一样）、句式可以不依赖动词而指派论元等论断的可靠性就要大打折扣了。

3.2 语义场景和基本层次概念

关于句式对动词的选择限制，Goldberg (1995)指出：句式必须指定动词跟它们结合的方式、限定可以通过各种方式跟它们整合的动词类别、指定动词所表示的事件类型整合进句式所表示的事件类型的方式，这就是动词与句式整合的原则（第 49 页）。那么，什么样的动词可以进入什么样的句式呢？Goldberg (1995)指出：动词所指的事件类型是句式所指的更为一般的事件类型的实例。……不包含直接跟句式相关的意义的动词经常指一种实施这种行为的方式（第 60 页）。用这种标准来衡量§3.1 中的例(1)—(13)，那么我们就会产生疑惑：为什么 give、tell、bake、threw、permit、allow、refuse、deny 可以作为双及物句式表示的各种转让意义的实例，而意义相似的 donate、whisper、ice、blast、let、enable、prevente、disallow、forbid 却不行？为什么"吃、穿、戴"可以作为双宾语句式表示的各种取得意义的实例，而意义相似的"啃、嚼、尝、披、围"却不行？为什么"飞"可以作为"NP (E)＋V＋NP (Th)"句式所表示的丧失意义的实例（具体地指示了丧失的方式），而意义相似的"蹿、蹦、跳、溜、走、滚、爬、游"却不行？

根据上文§2.2 的讨论，不包含跟某种句式直接相关的意义的动词（简称边缘动词），是通过套用这种句式、代换包含跟该句式直接

相关的意义的动词(即典型动词)而进入这种句式的。边缘动词的意义必须可以解释为典型动词的意义的一个次类,前者具体地例示(instantiate)后者的手段(means)、方式(manner)、条件(precondition)、结果(result)等,从而使语言表达更加精细化。比如,bake、threw、permit、allow、refuse、deny 等说明了给予(或不给予)的具体的方式或条件,"吃、穿、戴"说明了取得的手段或结果,"飞"说明了丧失的方式。但是,受动词意义必须跟句式意义相协调的原则的制约,这种精细化表达是有一定的限度的;具体地表现为:句式只能容忍在概念层级上比典型动词低一个级别的边缘动词、而不能容忍比典型动词低两个、或更多级别的边缘动词。比如,tell、threw 等可以看作是 give 的低一个级别的实例,而 whisper、blast 则是更为下位的方式动词;"吃、穿、戴"等可以看作是"拿"等取得意义的下位动词,而"啃、嚼、尝、披、围"则是"吃、穿、戴"等的下位动词,表示更加具体的方式或手段。如果引入 Lakoff (1987)中关于基本层次范畴(basic-level categories)的概念,那么我们可以发现:能替换典型动词进入某种句式的边缘动词必须是表示基本层次概念的。像上面的"吃、穿、戴"等是表示基本层次概念的,而"啃、嚼、尝、披、围"则是表示比基本层次概念更为具体和下位的概念的。再如:

(1) a. 我扔小明一个高抛球
b. 我踢小明一个斜线球
c. *我磕小明一个斜线球
d. *我顶小明一个斜线球
e. *我甩小明一个斜线球
f. *我钩小明一个斜线球
g. *我铲小明一个斜线球

(2) a. 我传小明一个高抛球
b. *我托小明一个高抛球
c. *我垫小明一个高抛球
d. *我推小明一个斜线球
e. *我扣小明一个斜线球

例(1)中的动词都是用于足球运动这种场景的,对于传送足球这种动作来说,"扔、踢"是表示基本层次概念的,而"磕、顶、甩、钩、铲"则是更加具体和专门的动作。例(2)中的动词都是用于排球运动这种场景的,对于传送排球这种动作来说,"传"是表示基本层次概念的,而"托、垫、推、扣"则是更加具体和专门的动作。

因为一个句式只能表示一个场景(scene),句子所表示的语义场景作为一种理想化的、内部一致的、个别性的行为或过程,[1]它通常是由典型的、容易激活这种情景的动词来表达的。特别是当句式通过引申用法而接纳边缘动词来充当谓语核心时,要求动词所传达的意义尽可能地接近典型动词,至少可以解释为是典型动词的直接的下位概念(比如,表示了典型动词所表示的动作行为的具体的方式)。一般地说,这种表示了某种上位动作和行为的动词是基本层次的概念,表示更为具体和专门的动作和行为的方式的动词一般是非基本层次的概念。

3.3 义项固定、词汇衍生和论元结构改变

对于某种句式来说,边缘动词的意义跟这种句式的意义是有一定差距的。为了让动词更好地适合句式意义,特别是为了让动词的参与角色能跟句式的论元角色相熔合(fusion);有一种词汇化(lexicalization)的办法可以使边缘动词逐渐逼近并成为典型动词,那就是:在句式意义的强力渗透和典型动词的同化(assimilation)作用下,边缘动词本身引申出跟句式意义相吻合的新的义项,或者说是句式意义部分地积淀和固化到词项意义上。例如:[2]

(1) a. Pauling smiled.
(鲍玲露出了微笑)
b. Pauling smiled her thanks/approval.
(鲍玲以微笑表示谢意/同意)

---

① Fillmore (1977: 84)对场景的定义是:一个理想化的、内部一致的、个别性的感觉、记忆、经验、行为或事物。

② 例子和释义,分别参考 Goldberg (1995);《牛津高级英汉双解词典》(第四版增补本),商务印书馆,2002 年;《新英汉词典》(增补本),上海译文出版社,1985 年。

(2) a. My father frowned.
(我父亲皱眉头了)
b. My parents always frown on late night out.
(我父母向来不赞成深夜外出)
c. My father frowned away the compliment and the insult.
(我父亲用皱眉头来击退阿谀和冒犯)

(3) a. Bees are swarming in the garden.
(蜜蜂在花园里成群地飞)
b. The crowd was swarming out through the gate.
(人群一窝蜂地从大门涌出)
c. crowds swarming in the streets
(街上拥挤不堪的人群)
d. The garden is swarming with bees.
(花园里到处飞满了蜜蜂)

(4) a. Thunder is rumbling in the distance.
(远处的雷声隆隆作响)
I am so hungry that my stomach's rumbling.
(我饿得肚子咕咕叫)
b. The trams are rumbling through the streets.
(电车发着辘辘声驰过大街)
The truck rumbled down the street.
(卡车发出辘辘声驰过大街)

(5) a. The flies are buzzing round a pot of jam.
(苍蝇围着果酱罐头嗡嗡叫)
b. The fly buzzed into the room.
(那只苍蝇嗡嗡叫着〔飞〕进房间)

(1a)中不及物的 smile 本来指微笑(give a smile),这是一种用以表示幸福、快乐、满足的行为和表情;扩大到用微笑来表示某种信息(express sth by means of a smile),引申出(1b)这种及物动词的意义和用法。同样地,(2a)中不及物的 frown 本来指皱眉,这是一种用以表示生气、沉思、忧愁的行为和表情,扩大到用皱眉来表示不赞成,引

申出(2b)这种及物动词的意义和用法;再引申一步,很容易引申出(2c)这种指用皱眉来做某事的意义,尽管一般的辞书还没有收录这个义项。(3a)中不及物的 swarm 本来指(蜜蜂)成群飞行,引申指(3b)所示的成群地移动和(3c)所示的聚集,最后引申指(某处)挤满了(人或物)。(4a)中的 rumble 本来指发出持续的低沉的声音,引申指(4b)所示的发出低沉的声音(沿着某个方向)行进;相似地,(5a)中的 buzz 本来指发出嗡嗡的声音,用在(5b)这样的句式中,很容易引申出指发出嗡嗡的声音(沿着某个方向)行进这种意义,尽管一般的辞书还没有收录这个义项。这种增加义项的办法主要针对个别语义有特异性的词汇。新的义项带来新的跟句式更加吻合的论元结构。

针对成批的有句法、语义共性的动词小类,可以通过词汇衍生(lexical derivation)手段,在不改变词义的情况下改变动词原有的论元结构,从而创造出适合某种句式的某种类型的动词或动词性结构的特有的论元结构。例如:

(6) a. 门口蹲着一个小孩

　　b. *门口哭着一个小孩

(7) a. 身后站着一个卫兵

　　b. *身后笑着一个卫兵

(8) a. 桌子上放着一本词典

　　b. *桌子上做着一个蛋糕

(9) a. 墙上画着一幅山水画

　　b. *床上脱着一双红袜子

(10) a. 小明在桌子上放了一本词典

　　b. 桌子上被小明放了一本词典

　　c. 桌子上小明放了一本词典

　　d. 桌子上放了一本词典

(11) a. 老张在墙上画了一幅山水画

　　b. 墙上被老张画了一幅山水画

　　c. 墙上老张画了一幅山水画

　　d. 墙上画了一幅山水画

从例(6)(7)来看,同样是一价动词,为什么"蹲、站"可以进入"NL+V着+NP"句式,而"哭、笑"却不能?从例(8)(9)来看,同样是二价动词,为什么"放、画"可以进入"NL+V着+NP"句式,而"做、脱"却不能?从句式语法的角度,可以这样回答:因为存在句式"NL+V着+NP"表示一种存在状态,要求其中的动词必须是包含〔状态〕、〔附着〕意义的定位(placement)动词。① 也就是说,"蹲、站、放、画"在语义上都隐含着一个处所论元;因此,"蹲、站"实际上是能支配客体(theme)和处所(location)两个内在角色(intrinsic role)的二元动词,"放、画"实际上是能支配施事、客体和处所三个内在角色的三元动词。

令人感兴趣的问题是,在存在句"NL+V着+NP"中,为什么不能出现施事论元?一种办法是,假设"放、画"等定位动词有两种论元结构:一种有施事论元,如例(10a-c)和(11a-c)所示;一种没有施事论元,如例(8a)和(9a)所示。但是,这类动词是大量的,这种增加义项的做法会大大地增加说话人大脑中心理词典(mental dictionary)的负担;也不符合儿童语言习得的实际情况,没有证据表明儿童把(8—9)和(10—11)中的"放、画"等当作两种义项来学习的。为此,Pan (1996) 提出了一条通用的词汇规则——施事删除(agent deletion)规则。即非完成体标记"着"附着在动词之后,可以把施事论元删除;"着"引发施事删除的条件是:(i) 相关动词是定位动词,具有{施事,客体,处所}三种论元角色;(ii) 客体和处所有一种像主语和谓词一样的关系,即处所是客体所在的地方。② 因此,下面这种句子是不合格的:

(12) a. *桌子上小明放着一本词典

b. *桌子上被小明放着一本词典

(13) a. *墙上老张画着一幅山水画

① 关于这种动词的语义特征和句法表现,详见朱德熙(1981/1990)等著作。

② 为了生成合格的处所倒装句(location inversion sentence),潘海华(1997:10)修正了词汇映射理论(lexical mapping theory)中的特殊默认分类(special default classification):赋予施事(它是可有可无的)一种〔+受限制〕的特征(因此只能作由介词引导的间接格或语义上受到限制的宾语),赋予客体一种〔+焦点〕的特征(因此只能居于动词后面宾语的位置),赋予处所一种〔(受限制〕的特征(因此可以作主语或宾语)。于是,对于由"着"引发的删除了施事的动词来说,其客体论元只能作宾语,其处所论元只能作主语(因为宾语位置已经被客体占领了)。

b. *墙上被老张画着一幅山水画

显然，施事删除规则无法推广到例(10b—d)和(11b—d)这种动词带"了"的句子上。因为，在这种句子中，施事可以不出现，如例(10d)和(11d)所示；但也可以出现，如例(10b—c)和(11b—c)所示。为此，我们吸收顾阳(1997/1999)和潘海华(1997)的若干思想，作出如下假设：在例(10b—c)和(11b—c)中，动词的论元结构经历了另一种词汇规则的作用，那就是广义被动化(generalized passivization)规则。[①] 在这种广义被动化规则的作用下，处所论元升级(promotion)了，表现为：不

① 顾阳(1997)采纳 Levin & Rappaport (1995)的理论假设：在词库(lexicon)和句法表达(syntactic representation)层面之间有两个界面：(i) 词汇语义表达式(lex ical-semantic representation)，(ii) 词汇句法表达式(lexical-syntactic representation)，也称为论元结构(argument structure)；词汇从词库到句法层面要先经过词汇语义表达式，再经过词汇句法表达式。某些词汇经过这两个层面可以衍变为新的词汇，如非宾格动词(unaccusative verb)、中间动词(middle verb)等。并认为，"着"规则(即施事删除规则)作用于词汇语义表达层面；由于词汇语义表达式和句法层面之间隔了一个词汇句法表达式，因而被控制的施事在句法表达层面上是绝对反映不出来的。而被动化规则作用于词汇句法表达式(论元结构)层面，即在词汇句法表达式(论元结构)层面上施事论元受到控制；由于词汇句法表达式和句法表达层面之间不存在其他表达式，因而受控制的施事在句法层面上仍然可以表现出来(第 23 页)。但是，她没有提到"桌子上小明放了一本词典、墙上老张画了一幅山水画"这类句子；因此，我们不知道在她心目中这种句子是主动式还是被动式。潘海华(1997)采用的是词汇映射理论(lexical mapping theory，简称 LMT)，相信在词库和句法表达层面之间只有一个层次，那就是论元结构。因此，他不利用层次的概念，而只是规定施事删除规则和被动化规则的操作结果不同："着"规则确实把施事给删除了，而被动化规则只是把施事降级了(第 12—13 页)。另外，他认为带"了"的存现句是有多种来源的。例如：

(1) a. 桌子上小明放了一本词典 → b. 桌子上放了一本词典

(2) a. 墙上老张画了一幅山水画 → b. 墙上画了一幅山水画

(3) a. 桌子上被小明放了一本词典 → b. 桌子上放了一本词典

(4) a. 墙上被老张画了一幅山水画 → b. 墙上画了一幅山水画

他把(1a)(2a)中的处所词语"桌子上、墙上"看作是话题，把其中的施事"小明、老张"看作是主语；而把(3a)(4a)中的处所词语"桌子上、墙上"看作是主语，其中的施事"小明、老张"看作是间接格。并且，他认为"桌子上放了一本词典、墙上画了一幅山水画"等存现句是有歧义的：它们既可以是从(1a)(2a)删除施事主语而得到的(1b)(2b)，也可以是从(3a)(4a)删除间接格施事而得到的(3b)(4b)(第 10—11 页)。

需要介词引导直接作句子的主语和话题;但是,施事论元被降级(demotion)了,表现为:(i) 或者用介词"被"引导,居于修饰语(状语)的位置,如例(10b)和(11b)所示;(ii) 或者不用介词"被"引导,居于内层主语(小主语)的位置,如例(10c)和(11c)所示;(iii) 或者干脆省略掉,如例(10d)和(11d)所示。"了"引发施事降级的条件是:(i) 相关动词是定位动词,具有{施事,客体,处所}三种论元角色;(ii) 客体论元和处所论元有一种像主语和谓词一样的关系,即处所是客体所在的地方;(iii) 客体论元是句子的焦点(focus),处所论元是已知信息(given information)。通过这种在词汇规则控制下的词汇衍生过程,在不改变词义的情况下得到了跟句式的论元结构相匹配的动词的论元结构。显然,这是一种比增加义项要经济得多的手段;并且,对于语言学习者来说也有相当的可学性(learnability)。

4 句式扩张的认知基础和逻辑机制

4.1 句式套用的认知基础:隐喻投射和完形包装

上文(§2.2 和 §2.3)指出,句式套用和动词代入不仅使得动词的配价跟句式配价不一致,而且还造成了角色错配,即动词的论元结构跟句式的论元结构的不一致,并且是句式的论元结构压倒了动词的论元结构。现在的问题是:句式套用的认知基础是什么?换句话说,当甲类动词套用乙类动词的习用句式时,说话人在其概念结构中到底做了些什么工作呢?一个简单的回答是隐喻投射(metaphor projection),即把跟乙类动词及其习用句式相关的概念结构投射到甲类动词上,从而把甲类动词所表示的事件纳入乙类动词及其习用句式所表示的事件图式(event scheme)中。例如:

(1) a. 大张 扔 小刘 一包香烟
　　← a'. 大张 给 小张 一包香烟
　　b. 小平 灌 李伟 一杯白酒
　　← b'. 小平 给 李伟 一杯白酒

c. 小明 踢 小华 一个斜线球

← c′. 小明 给 小华 一个斜线球

d. 玉芳 孝敬 公公 一条香烟

← d′. 玉芳 给 公公 一条香烟

(2) a. 李铎 吃了 小邵 一个苹果

← e′. 李铎 拿了 小邵 一个苹果

b. 玲玲 只穿过 姥姥 一件毛衣

← f′. 玲玲 只拿过 姥姥 一件毛衣

c. 老刘 抽了 小孙 一支香烟

← h′. 老刘 拿了 小孙 一支香烟

d. 小芳 花了 奶奶 一百块钱

← g′. 小芳 拿了 奶奶 一百块钱

e. 小明 糟蹋了 我 好几张宣纸

← i′. 小明 拿了我 好几张宣纸

f. 王平 坑了 爸爸 一千块钱

← j′. 王平 拿了爸爸 一千块钱

(3) a. 王冕 死了 父亲

← a′. 王冕 失去了 父亲

b. 王大爷 飞了 一只鸽子

← b′. 王大爷 失去了 一只鸽子

c. 老王 烂了几个橘子

← c′. 老王 失去了 几个橘子

d. 我家 报废了一台电视

← d′. 我家 失去了 一台电视

e. 我家 被偷了 一台电视

← e′. 我家 失去了一台电视

(4) a. 床上 躺着 一个病人

← a′. 床上 有一个病人

b. 楼上 住着 几个留学生

← b′. 楼上 有 几个留学生

c. 园子里 种了 两棵枣树

← c′. 园子里 有 两棵枣树

d. 墙上 挂了 一幅山水画

← d′. 墙上 有 一幅山水画

在例(1)中,把双及物动词"给"及其习用的双宾句式所表示的"给予"性转让的概念结构投射到"扔、灌、踢、孝敬"等动词所表示的事件上,从而把"扔、灌、踢、孝敬"等单及物动词所表示的事件纳入双宾句式之中,使扔、灌、踢、孝敬等行为成为给予行为的一种具体的方式。在例(2)中,把双及物动词"拿"及其习用的双宾句式所表示的"获取"性转让的概念结构投射到"吃、穿、抽、花、糟蹋、坑"等单及物动词所表示的事件上,从而把"吃、穿、抽、花、糟蹋、坑"等动词所表示的事件纳入双宾句式之中,使吃、穿、抽、花、糟蹋、坑等行为成为获取行为的一种具体的方式。在例(3)中,把及物动词"失去"及其习用的主动宾句式所表示的"消失"性受损的概念结构投射到"死、飞、烂、报废、被偷"等动词或动词性结构所表示的事件上,从而把"死、飞、烂、报废、被偷"等动词或动词性结构所表示的事件纳入主动宾句式之中,使死、飞、烂、报废、被偷等行为成为损失行为的一种具体的方式。在例(4)中,把存在动词"有"及其习用的处所性存在句式所表示的处所性存在的概念结构投射到"躺着、住着、种了、挂了"等动词性结构所表示的事件上,从而把"躺着、住着、种了、挂了"等动词性结构所表示的事件纳入存在句式之中,使躺着、住着、种了、挂了等状态成为处所性存在的一种具体的方式。这正好体现了句式语法关于句式语义和词项语义互动(interaction of construction meaning and lexical meaning)的观念:句式提供了结构上及语义上的基本框架,各个词汇成分根据其词类功能而填入句式框架的各种位置,并对整个句子的语义作出贡献。[①] 在这里,是谓语动词的语义使得给予、获取、丧失、存在等句式意义增加了方式的意义。

---

① 参考黄居仁等(1999)的有关讨论,第427—428页。

从上面的讨论可以看出，在把乙类动词的概念结构通过隐喻来投射到甲类动词的概念结构上的同时，甲类动词的概念结构被整合进了乙类动词的概念结构；于是，甲类动词在套用乙类动词的惯用句式的同时，不仅获得了乙类动词的这种惯用句式的句式意义，而且还增加了由甲类动词所带入的意义。比如，例(1a—d)不仅表示给予，还指示了具体的给予方式；例(2a—f)不仅表示获取，还指示了具体的获取方式；例(3a—e)不仅表示损失，还指示了具体的损失的方式；例(4a—d)不仅表示存在，还指示了具体的存在方式。换句话说，句式套用的语义后果是把两种概念结构整合成一种新的复合性的概念结构，形成一种新的认知图式。或者说，把两种事件结构整合进一个完形(Gestalt)中，用一个认知图式来包装一个复合事件。例如：

(5) a. 一个月的工资全被他喝了猫儿尿了

← b. 一个月的工资全被他花了

在(5a)中，花钱和喝酒两个事件被整合进一个心理图式中，即用一个完形来包装。

但是，句式对事件结构的完形包装是有一定的限度的。一个语言可以选择某些(操该语言的人们认为)在认知上重要的事件、经验和知识用某种(或某些)句式来表达。至于选择哪些事件、经验和知识，这些知识如何通则化(generalize)，则并无固定不变的程式可循。这就造成了句式跟所表达的事件之间的对应关系是有理可循的(即有理据的，motivated)，但是选择何种对应又不是某种固定形式的规律所能预测的(即是任意的，arbitrary)。[①] 例如：

(6) a. 张大爷飞了一只鹦鹉

b. ？张大爷飞了一只风筝

c. *张大爷飞了一个气球

d. *张大爷飘了一个气球

① 参考黄居仁等(1999)，第415—416页。

为什么张大爷丧失了一只鹦鹉可以套用句式“NP (E)＋V＋NP (Th)”，说成“张大爷飞了一只鹦鹉”；这是可以解释的：因为张大爷是损失事件的经历者、一只鹦鹉是所损失的客体、而飞的行为又是丧失行为的一个实例(客体鹦鹉通过飞走的方式使经历者受到损失)。但是，为什么风筝、气球通过飞走、飘走的方式使张大爷受到损失就不能套用句式“NP (E)＋V＋NP (Th)”，这就不容易说出一个令人信服的解释来。正是这种事件结构和句式包装之间任意性的对应关系，造成了§3.1 所说的句式的不完全能产性，即语义性质相似的动词不一定都能进入某种句式。

4.2 归纳和类推：超越动词配价和句式构造之间的循环论证

基于词汇主义立场的动词的配价或论元结构研究，受到猛烈批评的一个理由是它陷于循环论证(circularity)。Goldberg (1995：11)举了下面的例子来说明这一点：

(1) The horse kicks.
(2) Pat kicked the wall.
(3) Pat kicked at the football.
(4) Pat kicked Bob black and blue.
(5) Pat kicked the football into the stadium.
(6) Pat kicked Bob the football.

在(1)中，kick 是一元动词，因为它带了一个补足语；在(2—3)中，kick 是二元动词，因为它带了两个补足语；在(4—6)中，kick 是三元动词，因为它带了三个补足语。这等于是说：断定 kick 有可以带 n 种论元的意义(n-argument sense)是基于它可以跟 n 种补足语共现这种事实，而同时又声称 kick 可以跟 n 种补足语共现是因为它有可以带 n 种论元的意义。这就造成了循环论证。张伯江(1999：183)和沈家煊(2000：292)也出于这种对循环论证的顾忌，转而强调句式配价比动词配价更重要，只有树立句式配价的观念才能避免循环论证和“词无定价、离句无价”的厄运。

现在，我们要问的问题是：句式配价或句式的论元结构的理论能够逃脱循环论证的厄运吗？答案是不可能。比如，如果问为什么“他扔我一个球”属于三价句式，那么回答：因为它跟“他送我一本书”一样有施事、受事和与事三个论元；如果问为什么二元动词“扔”在句子“他扔我一个球”中可以跟施事、受事和与事三个论元共现，那么回答：因为“他扔我一个球”是三价句式。可见，当沈家煊(2000：293)按照Goldberg (1995)的思路，把配价看作是句式的属性，将句式配价定义为指抽象的句式配备的、与谓语动词同现的名词性成分的数目和类属(指施事、受事、与事、工具)时；就注定了要卷入句式的配价数目由句式中的论元数目来决定、句式中的论元数目由句式的配价数目来解释的循环圈，从而使得用句式配价来更好地说明动词跟相关名词性成分在组配上的合格性的目标落空。

其实，根据我们的想法，只要找到一个合适的逻辑起点，那么上述循环论证都是可以避免的。比如，遵循Bloomfield (1933：20)“分析语言时，只有归纳的概括才是有用的概括”的思想(中译本第21页)，袁毓林(1987/1993：171)指出：“向”(即价)是动词跟名词性成分发生句法、语义联系而表现出来的一种性质，它表征着动词在一个句法结构中所能关联的名词性成分的数量。因此，“向”是动词的组合功能的数量化：能和一个名词性成分发生主谓或述宾关系的动词叫单向动词，能和两个名词性成分发生主谓或述宾关系的动词叫双向动词，能和三个名词性成分发生主谓或述宾关系的动词是三向动词。“向”的基础是动词在句法结构中跟名词性成分发生组合关系的潜能，“向”是一种建立在句法基础上的语法范畴，是动词的组合功能的数量表征。袁毓林(1998)进一步指出：因为价反映了动词对其他词项的支配能力，具有不同的支配能力的动词有不同的价；这样，通过对不同的动词的价的描写就可以对它们的句法组合能力作出简洁的刻画。也就是说，价反映了动词的某种分布状况——它到底能跟多少、哪些从属成分共现；或者说，价是对某种分布的集约化的表示——用数字来反映动词能跟多少从属成分共现。如此看来，研究配价的目的在于更好地说明

句法结构的合格性、说明句法结构跟语义结构的关系（第87页）。在这样的认识指导下，我们从一定数量的实际语料中归纳、总结各种动词的配价情况，概括出它们的论元结构；然后预测，在其他语境下，这些动词跟名词性成分的组配情况将是什么情况。这就像是词类划分一样，根据一定数量的实际语料，归纳、总结词的各种分布情况，把不同分布的词划归不同的词类；然后预测，这些词将各有什么样的分布位置和使用方式。从方法论上讲，这是一种基于用法的语法模型（usage-based modal of grammar），[①]从归纳中得到一般性的概括，再用一般性的概括来解释已有的相关现象并预测可能出现的相关现象。

上面这种思想可以得到语言习得方面的证据的支持，Goldberg（1995）指出：说话人在使用词汇时倾向于保守。人们通常把词汇使用于同样的句式中，他们以前听到过这些词项被别人用到这些句式中。但是，如果被适当地启动，他们也会把这种用法扩展到新的模型上。……新的用法和意义是通过跟既有的例子的相似性而获得的，……动词的小类是由说话人内在地、隐含性地对学过的例子进行概括而得出的。因为记忆是联想性的，用在同样句式中的类似的动词通过一般的范畴化过程而划归到一类中去（p. 133—4）。总而言之，通过归纳和类推，人们可以获得关于动词和句式之间互动关系的全部知识。

4.3　句式扩张的逻辑机制：归因推理和动因解释

根据上文的讨论，每一种句式都有一组惯用的动词，当其他类别的动词代入这种动词所惯用的句式时就造成了句式套用，句式套用的一个重要的语用动因是表达的精细化。同时，句式套用的一个直接的语法后果是句式扩张，这至少包括句式意义的引申和进入句式的动词类别的增加两个方面。那么，句式扩张的逻辑机制是什么呢？特别是§4.1中所说的：句式跟所表达的事件之间的对应关系是有理据的，但又是不可预测的，这种情况到底是否符合人类的思维规

① 基于用法的语法模型，详见Goldberg（1995），p. 133—139，192，226。

律？或者说人脑能否处理这种扑朔迷离的现象？对此，我们尝试从归因推理和动因解释的角度作出一些说明。

所谓归因(abduction)，就是推出最好的解释。为了进行归因，人们必须先知道结果，因此，归因推理涉及事后推理(after-the-fact reasoning)，用以决定为什么一连串特定顺序的事件是按照这种顺序发生的。可见，归因推理试图推出一连串已经发生的事件之所以是这种发生顺序的动因，但是它不能事先预测这一连串事件的发生顺序。这跟演绎(deduction)不同，演绎推理追求对一连串事件的发生顺序作出预测。现在，人工智能研究领域越来越清楚地认识到寻找动因之类的推理(motivation-like reasoning)的重要性，因为人类的许多智能行为就是基于通过寻找动因之类的推理来推出最好的解释。因此，归因对于建立人类自然的推理模式是有用的。这一点对于以模拟人类自然智能为目标的人工智能研究来说，意义十分重大。有研究表明，在语言运用方面人们广泛地使用归因推理。比如，尽管说话人不能预测两个相关的概念是否、或多大程度上会在形式上也相关，但是，为了使这种输入形式有意义、从而把这种新形式放入相关格式组成的网络(这构成了他们的语言知识)中，他们还是要寻找这种关系。也就是说，形式与意义、形式一意义配对之间的关系被语言使用者(无意识地)按照他们自己的方式观察和思考。显然，如果这种说法是正确的：人们寻找归因解释(即动因)来解释事件的顺序，那么，我们有理由猜想：说话人也许无意识地应用同样的原理来习得语言。① 对于特定句式及其句式意义和构成成分及其词汇意义之间的关系来说，语言使用者倾向于先验地认为：一定的句式表示一定的句式意义，一定的句式意义又是跟一定句式的特定构造相关的；句式中各个构成成分及其结构关系对句式的整体意义都有贡献，一定的构成成分由

---

① 以上叙述主要按照 Goldberg (1995)，p. 71，但也根据笔者的知识和理解作了引申和发挥。把 abduction 译作"归因"，是采纳了美国 Temple University 计算机系王培教授的意见，他在第三届国际认知科学大会(2001 年 8 月 27—31 日，北京)期间，鼓动我用归因推理等非公理逻辑(Non-Axiomatic Logic)的办法处理自然语言，谨此致谢。

一定的词汇或语法范畴来实现（或者说句式中的特定位置由特定的词汇、语法范畴来填充）。这样，就把句式意义归结为特定句式的整体构造和构成成分及其结构关系。例如：

(1) Pat handed Chris the ball.
(2) Pat threw Chris the ball.
(3) Pat hit Chris the ball.
(4) Pat shined Chris the ball.

人们从例(1)这样的句子上推出：双宾句的谓语动词要表示传递意义，以便能够联结施事、受事和与事三个论元。于是，对于例(2)(3)这样的句子也乐意接受；因为动词 threw 和 hit 的意义不仅跟 hand 等动词的意义不相冲突，并且可以解释为传递意义的下位概念——即具体地指示了转递的方式。更进一步，居然还能接受例(4)这样的句子；当然，脑筋得多转几个弯：先假定例(4)是合格的，并且表达了类似例(1)—(3)这种受事由施事向与事转移的传递意义；再在传递这种句式意义和谓语核心 shin(胫)的词汇意义之间进行互动，假定名词 shin 所表示的肢体意义不仅跟传递这种句式意义不矛盾，而且能够整合(integrate)进这种句式意义之中；于是，把在双宾句的谓语核心位置上的 shin 的意义解释为通过胫的动作来传递；最终，达到了句子的形式和句子的意义、句子整体意义和句子成分的意义的互相协调和互相可以解释，即是有理可据的、具有动因的。结果，使得双宾句的句式意义得到进一步的引申和扩大，从原来表示单纯的传递到后来表示通过某种特定的方式来传递，一直到表示通过某个特定肢体的动作来传递。

更有甚者，人们还愿意把嵌在特定句式中的生造出来的词也解释为具有跟句式意义相协调的词汇意义。例如：

(5) She gave him something.
(6) She topamased him something.

Goldberg (1995：35)指出，他的十个被试中，竟有六个人认为无意义单词(nonsense word) topamase 的意思是 give。

总而言之，在寻找动因和最好的解释这种归因推理的逻辑机

制的作用下，填入特定句式的特定位置的词汇类别增加了，随之而来的是该句式的句式意义引申和扩大了。据此，归因推理可以看作是句式和词汇互动的一种逻辑机制。归因推理这种寻找动因解释但无法作出预测的逻辑机制，正好适合处理句式跟所表达的事件之间的对应关系是有理据的、但又是不可预测的这种语言现象。换句话说，句式跟所表达的事件之间的对应关系是有理据的、但又是不可预测的这种语言现象，是我们人脑的思维规律所允许的，也是能处理的。

5　论元结构和句式结构的互动：从原理走向规则

现在，大概多数语法研究者都能同意：句式意义主要来源于动词的论元结构和句式结构的交互作用(interaction)；但是，词汇与句式的交互作用只是一个基本的原理，我们应该把这种抽象的原理具体化为可以落实到具体的操作上的规则。比如，我们上文多次讨论到下列几种动词变价或论元增生的句式：

(1) a. 王冕 死了 父亲

b. 王大爷 飞了 一只鸽子

c. 老王 烂了 几个橘子

d. 我家 报废了 一台电视

(2) a. 李四 被后边的司机 按了一喇叭

b. 老师 被学生 贴了大字报

关于这种句子中增加的动词原来的论元结构中所没有的受害者论旨角色(maleficiary role)，潘海华(1997)认为：在汉语语法系统中，有一条普遍的受害者插入规则(general maleficiary role insertion rule，简称 MRI)，引发了把受害者角色加入相关动词的论元结构中的操作。控制这条规则操作的语义条件是动词的意义，这种动词要求带

有某种不好的效果或影响(第 6 页、第 15 页注 7)。[①] 我们认为,受害者插入规则可以推广为更加普遍的与事插入规则(general dative role insertion rule,简称 DRI),从而把§1.2 中的受惠者论元增容规则和目标论元增容规则也概括进来。控制这条与事插入规则操作的语法条件是句式的意义:当句式义涉及当事(包括受害者)和客体两个论元角色之间的丧失关系,而相关动词只有客体一个参与角色时;或者,当句式义涉及施事、受事和与事(包括受惠者、受害者和目标等)三个论元角色之间的转移关系,而相关动词只有施事和受事两个参与角色时;就启动与事插入规则,在相关动词的论元结构中插入了一个与事论元。比如,像例(1)这样的丧失句式要求有丧失的主体(即受害者)和丧失的客体两个论元角色,而"死、飞、烂、报废"这种动词原有的论元结构中只有一个客体论元;于是,在丧失句式的丧失意义的驱动下,在相关动词的论元结构中临时插入了一个受害者角色。同样,"被"字句通常表示不如意的遭受等意义;当述宾结构作"被"字句的谓语核心时,在遭受这种句式意义的驱动下,在相关动词的论元结构中临时插入了一个受害者角色。再如:

---

① 潘海华(1997)指出,带有受害者插入的动词(或者说得广一点儿,所有含有论旨角色"受害者"的动词)都只允许受害者作主语,而另一个论旨角色"客体"或"受事"则只能成为所谓的滞留宾语(retained object)。这种现象也包括下面这类句子:

(1) 他被我罚了五块钱

(2) 他被我踢了一脚

(3) 那块肉被我炒了青椒

他在词汇映射理论的框架内,利用受害者和客体/受事在论旨层级关系(thematic role hierarchy,简称 TRH)上的不同位置(施事>受益者/受害者>接受者/经验者>工具>客体/受事>处所),加上主语条件及其相关的映射规则来正确地预期受害者、而不是客体/受事,作被动句的主语。具体的技术细节请看潘海华(1997),第 6—7 页。但是,如果考虑到下列例子,那么潘海华(1997)"只允许受害者作主语"的论断就必须重新检讨:

(4) 李铎 吃了 小邵 一个苹果

(5) 老刘 抽了 小孙 一支香烟

(6) 小芳 花了 奶奶 一百块钱

如果其中的与事论元的论旨角色也是受害者,那么受害者角色也可以充当间接宾语。

(3) a. 李铎 吃了 小邵 一个苹果
b. 玲玲 只穿过 姥姥 一件毛衣
c. 老刘 抽了 小孙 一支香烟
d. 小芳 花了 奶奶 一百块钱
e. 小明 糟蹋了 我 好几张宣纸
(4) a. 王刚 扔 我 一包香烟
b. 小平 塞 李伟 一个纸条
c. 小明 踢 小华 一个斜线球
d. 玉芳 孝敬 公公 一条香烟

单及物动词"吃、扔"原有的论元结构中只有施事和受事两个论元，但是，在双及物句式的获得/给予意义的驱动下，在相关动词的论元结构中临时插入了一个与事（受害者或受益者）角色。这样，使得动词的论元结构跟句式的论元结构能够更加吻合。

我们希望能够发现更多的诸如普遍的与事角色插入规则、§3.3中讨论的施事删除规则和广义被动化规则等改变动词的论元结构的规则，并刻画控制这种规则使用的句法、语义条件，从而使词汇与句式交互作用的原理更加具体化和可操作化。

6　结语：一个简短的总结

§1指出动词配价学说和论元结构理论的核心思想是词汇主义和动词中心论，当碰到动词的组配性质跟句式构造不一致时，它们分别采用变价和论元增容规则来处理。与此相对的是句式语法和句式配价的思想，认为配价是句式的属性、句式具有指派论元的功能、句式的论元结构是由句式意义决定的。

§2指出基本的句式意义是由句式所惯用的典型动词的论元结构带来的，一定类别的动词有其惯用的句式。为了表达的精细化等语用动机，甲类动词可以有条件地套用乙类动词所惯用的句式；这造成动词的论元结构和句式的论元结构不一致，还造成了句式意义的引申和扩张。句式套用和词项代入是动词和句式互动的一种具体的

操作机制。

§3 指出句式对动词有严格的、难以预测的选择限制，这造成了句式的不完全能产性。只有当动词表示的事件类型是句式所表示的事件类型的一个直接的下位概念、并且这个下位概念是基本层次范畴时，才能代入这种句式。代入某种句式的非典型动词跟这种句式有两种磨合方式：一是通过词汇化手段增加新的义项，从而产生跟句式相一致的论元结构；二是通过词汇衍生手段，在不改变词义的情况下改变动词的论元结构。

§4 指出甲类动词套用乙类动词所惯用的句式的认知基础是隐喻投射，即把跟乙类动词及其惯用句式所表示的概念结构投射到甲类动词所表示的事件结构上；从而把两种事件结构合并成一个事件结构，包装进一种事件图式中，并形成一个心理完形。以归纳动词的句法组配模式为逻辑起点，可以超越动词配价和句式构造之间的循环论证。用类推的观念，可以解释句式意义的引申和动词类别的扩展。指出追求动因解释的归因推理是句式扩张的逻辑机制，藉此可以解释句式与所表示的事件在对应关系上的有理据性和不可预测性。

§5 指出应该概括出诸如普遍的与事插入、施事删除、广义被动化等改变动词的论元结构的规则，从而使动词与句式互动的原理具体化和可操作化。

要而言之，表达的精细化等语用动机促动了句式套用和词项代入，这又引发了动词和句式的互动，其结果是动词改变其论元结构来适应句式意义和句式构造的需要。在一定的句法、语义条件下，与事插入、施事删除等规则作为动词和句式互动的具体机制，使动词衍生出符合句式要求的论元结构。动词和句式的对应关系是有理据的、但又是不可预测的，动词和句式互动背后的逻辑机制是追求动因解释的归因推理。

鸣谢：本文先后承顾阳、郭锐等学长指正，谨此致以诚挚的谢意。

参考文献

顾　阳 (1994)《论元结构理论介绍》,《国外语言学》第1期,第1—11页。

顾　阳 (1996)《生成语法及词库中动词的一些特性》,《国外语言学》第3期,第1—16页。

顾　阳 (1997)《关于存现结构的理论探讨》,《现代外语》第3期;略作修改后收入徐烈炯(1999),第91—110页。

顾　阳 (1999)《双宾语结构》,收入徐烈炯(1999),第60—90页。

黄居仁、张莉萍、安可思、陈超然 (1999)《词汇语意和句式语意的互动关系》,《中国境内语言暨语言学》第五辑:《语言中的互动》,第413—438页,台北:中研院语言学研究所筹备处。

陆俭明 (2002)《再谈"吃了他三个苹果"一类结构的性质》《中国语文》第4期,第317—325页。

马庆株 (1983)《现代汉语的双宾语构造》,《语言学论丛》第十辑,商务印书馆。收入马庆株(1992),本文据此。

马庆株 (1992)《汉语动词和动词性结构》,北京语言学院出版社。

马庆株 (1998)《动词的直接配价和间接配价》,收入袁毓林、郭锐(1998),第283—294页。

潘海华 (1997)《词汇映射理论在汉语句法研究中的应用》,《现代外语》第4期。

沈家煊 (1999)《"在"字句和"给"字句》,《中国语文》第2期,第94—102页。

沈家煊 (2000)《句式和配价》,《中国语文》第4期,第291—297页。

徐烈炯 (1999) 主编《共现与个性—汉语语言学中的争议》,北京:北京语言文化大学出版社。

袁毓林 (1987)《准双向动词研究》,杭州大学硕士论文,摘要发表于《语言研究》1989年第1期,全文作为附录收入袁毓林(1993)《现代汉语祈使句研究》,北京大学出版社。

袁毓林 (1998)《汉语动词的配价研究》,江西教育出版社。

袁毓林、郭　锐 (1998) 主编《现代汉语配价语法研究》第二辑,北京大学出版社。

张伯江 (1999)《现代汉语的双及物句式》,《中国语文》第3期。

朱德熙 (1978)《"的"字结构和判断句》,《中国语文》第1、2期。收入朱德熙(1980),125—150页,本文据此。

朱德熙 (1979)《与动词"给"相关的句法问题》,《方言》第2期。收入朱德熙

(1980),151—168 页,本文据此。

朱德熙 (1980)《现代汉语语法研究》,商务印书馆。

朱德熙 (1981)《"在黑板上写字"及相关句式》,《语言教学与研究》第 1 期。此文最早在该杂志的第三集(1978 年 5 月)上发表,这是修改稿。后来作了修改,收入朱德熙(1990)第 1—16 页,本文据此。

朱德熙 (1990)《语法丛稿》,上海教育出版社。

Chomsky, Noam (1981) *Lectures on Government and Binding*. Dordrecht: Foris.《支配和约束论集——比萨学术演讲》,周流溪、林书武、沈家煊译,赵世开校,中国社会科学出版社,1993 年。

Chomsky, Noam (1992) *A Minimalist Program for Linguistic Theory*. MIT Occasional Papers in Linguistics 1. Cambridge, Mass.: Dept. of Linguistics and Philosophy, MIT.

Bloomfield, Leonard (1933) *Language*. New York: Holt, Rinehart & Winston.《语言论》,袁家骅、赵世开、甘世福译,钱晋华校,商务印书馆,1985 年。

Dowty, D. (1991) Thematic Proto-role and Argument Selection. *Language*, Vol. 67, No. 3.

Fillmore, C. (1968) The Case for Case. *Universals in Linguistic Theory*, ed. by Emmon Bach and Robert T. Harms, 1—90. New York: Holt, Rinehart & Winston.《"格"辨》,胡明扬译,《语言学译丛》,第二辑,第 1—117 页,中国社会科学出版社,1980 年。

Fillmore, C. (1977a) The Case for Case Reopened, in P. Cole & J. M. Sadock (eds.) *Syntax and Semantics*, Vol. 8, *Grammatical Relations*, pp. 59—81. Academic Press.

Fillmore, C. (1977b) Topics in Lexical Semantics, in R. W. Cole (ed.) *Current Issues in Linguistic Theory*, pp. 76—138.

Goldberg, E. Adele (1995) *Constructions: A Construction Grammar Approach to Argument Structure*. Chicago and London: The University of Chicago Press.

Grimshaw, J. (1990) *Argument Structure*. MIT Press.

Gruber, J. (1976) *Lexical Structures in Syntax and Semantics*. North-Holland.

Hale, K. & S. J. Keyser (1991) *On the Syntax of Argument Structure*. MIT Press.

Jackendoff, Ray. (1972) *Semantic Interpretation in Generative Grammar*, MA: MIT Press.

Jackendoff, Ray (1990) *Semantic Structure*. The MIT Press, Cambridge, Massachusetts.

Lakoff, George (1987) *Women, Fire, and Dangerous Things: What Categories Reveals about the Mind*. Chicago & London: The University of Chicago Press.

Larson, Richard (1988) On the Double Object Construction. *Linguistic Inquiry* 19: 335—391.

Larson, Richard (1990) Double Objects Revisited: A Reply to Jackendoff. *Linguistic Inquiry* 21: 589—635.

Levin, Beth & Malka Rappaport (1995) *Unaccusativity: At the Syntax-Lexical Interface*. Cambridge: MIT Press.

Levin, Beth & Malka Rappaport (1997) Lexical Semantics and Syntactic Structure, in Lappin, Shalom (ed.) (1997) *The Handbook of Contemporary Semantic Theory*, 487—508, Oxford: Blackwell Publishers.

Pan, Haihua (1996) Imperfective Aspect *zhe*, Agent Deletion, and Locative Inversion in Mandarin Chinese. *Natural Language & Linguistic Theory*, 14: 409—432.

2002年10月初稿，2003年2月改定

（删节发表于《语言研究》2004年第4期）

三、信息抽取和语义标注

信息抽取的语义知识资源研究

本文讨论支持信息抽取的语义资源的建设问题，举例说明了信息抽取至少需要三种层面的语义知识：(i) 宏观的话语篇章知识，藉此可以约束信息抽取的匹配模板的类型，预测关键性的信息项目在文本中的分布位置；(ii) 中观的论元结构知识，藉此可以建立动词的论元成分跟事件模板的传递与继承关系，帮助确定代词或空语类跟其先行语的回指关系，进而确定其语义所指；(iii) 微观的逻辑结构知识，藉此可以确定否定词、量化词、模态词等逻辑算子跟其所约束的成分之间的逻辑关系（比如，哪些成分处于否定的辖域之中，其中哪个成分是否定的焦点，在哪些语法条件下否定词是冗余的，等等）。最后，指出研究这三种语义知识所可利用的几种理论和方法。

1 信息抽取和语义知识资源

信息抽取(information extraction，简称 IE)指用计算机自动地从一段文本(text)中抽取出指定的一类信息（比如，事件、事实等），并将其形成结构化的数据填入一个数据库中供用户查询和使用的过程。[①] 例如，从一篇关于军事演习的新闻报道中摘录出演习的类型、时间、地点、兵种、武器、装备、假想敌、后勤保障等信息。随着互联网的普及，信息抽取成了从网上自动地获取自然语言文本中特定信息的一种非常简捷和有效的途径；相应地，信息抽取技术成了计算机科学和计算语言学中的一门具有良好的发展前景的应用技术。目前国外从事信息抽取研究主要有基于知识和基于统计两种方法，我们认为单纯地用基于知识的模板匹配方法或者单纯地用基于数学的概率统计方法都是有缺陷的。可以预期，把这两种方法结合起来将是今后研究的必然趋势；但是，怎样把这两种方法

① 详见孙斌(2000)，第 104—105 页。

结合起来，一时还不容易找到可行的交叉点。我们从语言学的角度构想：如果对相当数量的真实文本进行语义关系标注，建立一种标有语义关系网络的精炼语料库（简称网库〔net-bank〕）作为信息抽取的资源，那么对于基于知识的方法来说，这种网库将提供比脚本和框架更为精细的语义关系网络，有利于对相关信息进行比较和甄别，从而提高信息抽取的精度。对于基于统计的方法来说，这种网库将提供一种语义信息十分丰富的训练样本，可以通过机器学习的方法来提高信息抽取的速度。

事实上，研制一种面向信息抽取的汉语真实文本的语义标注系统，不仅能为汉语文本的信息自动抽取提供可靠的基础和充分的资源，而且对于计算机信息检索、自动文摘、信息理解、文本相似度的自动比较乃至语音识别、“汉语-外语”之间的机器翻译等自然语言信息的自动处理，提供新的观念、方法和资源，从而为中文信息处理技术的产品化乃至产业化，从语言学角度作出新的贡献。

要研制这种网库，首先要对语义标注的理论和方法进行研究，特别是要研究和发现词项之间、句子之间什么样的语义关系对于信息抽取是有用的。这必将推动我们去检讨已有的各种语义分析的理论和方法的效用和不足，努力发掘新的、对于抽取信息有用的语义关系，发展新的语义分析的理论和方法，反过来从技术应用的角度对理论语言学作出贡献。

2 哪些语义知识对信息抽取是重要的

为了真正为信息抽取提供有用的语义知识方面的资源，首先必须对真实文本的不同层次的语义关系进行分析和辨别。围绕着信息抽取这种实用的目标，研究下列问题：(1) 什么样的语义知识对从真实文本中抽取出关键性的指定信息是有用的、并且是十分重要的；(2) 怎样在真实文本上标注这些语义信息，并且使计算机可以自动识别和作为训练样本来利用。由于篇幅的限制，本文只能先讨论第一个问题。

我们通过对相当数量的现代汉语真实文本的分析，发现以下这些语义知识对信息抽取是至关重要的：话语结构、篇章关系、论元结构、题元角色、角色转换、照应关系、否定结构、辖域歧义、算子约束关系等，它们可以归并为如下三大类：

2.1 宏观层次的话语—篇章知识(discourse-text knowledge)

以整个篇章为考察单位，包括文本中不同的段落和小节甚至句群和句子之间的语义关系。比如，哪些是话题句(topic sentence)或论点句，哪些是支持句(support sentence)或论证句，哪些是背景句(background sentence)，哪些是总结句(summary sentence)或结论句等；还有不同句子之间的逻辑语义关系，比如：条件、因果、转折等偏正关系或并列、对比、递进等联合关系等。因为这种信息对于指导计算机从文本的哪些地方、哪些句子中抽取信息有决定性的影响。比如，我们分析了几十篇关于领导人出访或会见、政府更迭等的通讯报道，发现新闻这种文体的信息分布是有很强的规律的；往往是标题或第一句话(即话题句)中差不多就包含了全部重要的信息项目。现在以新华通讯社 2001 年 4 月 27 日的《每日电讯》为例：

(1) 正标题：江泽民会见智利参议长

副标题：希望新世纪中智友好关系更上一层楼

第一句：新华社北京 4 月 26 日电(记者杨国强)国家主席江泽民今天下午在中南海会见智利参议长安德烈斯·萨尔迪瓦时说，智利是中国在拉美的重要合作伙伴。

(2) 正标题：李瑞环会见摩洛哥国王

副标题：指出中国人民不忘非洲“老朋友”永做非洲“好朋友”

第一句：新华社非斯(摩洛哥)4 月 26 日电(记者车玉明、范卫平)摩洛哥国王穆罕默德六世 26 日在非斯王宫和中国全国政

协主席李瑞环亲切会见。[①]

(3) 正标题：朱镕基将出访五国

第一句：新华社北京 4 月 26 日电外交部发言人章启月今天下午在记者招待会上宣布：应巴基斯坦……的邀请，朱镕基总理将于 5 月 11 日至 22 日对上述五国进行正式访问。

(4) 正标题：森内阁宣布总辞职 小泉新内阁组成

第一句：新华社北京 4 月 26 日电(记者王大军)日本森喜朗内阁 26 日上午在最后一次临时内阁会议上宣布总辞职，新的小泉新内阁将于当天晚间成立。

在新闻报道的标题或正文第一句(即话题句)中，一般首先要交代新闻的六要素，即：时间(when)、地点(where)、人物(who)、事件(what)、原因(why)和方式(how)，简称 6W，从而囊括了用户最关心的几个关键性的信息项目。可见，篇章结构对于关键性的信息项目的分布有很强的预示作用。

另外，文本的文体(style)类型方面的知识，对于信息抽取系统调用匹配模板的类型具有约束作用。比如，像新闻报道等叙事类(narrative)文体，一般要调用包含上述 6W 的模板；而像报刊社论、时事评论等议论类(argumentation)文体，一般要调用包含论点、论据、结论等的模板。

2.2　中观层次的论元结构知识(argument structure knowledge)

基本上以句子为考察单位，包括句子中的谓词(动词和形容词)和有价名词(一价名词和二价名词)跟其从属成分之间的支配和依存

① 这里"摩洛哥国王穆罕默德六世 26 日在非斯王宫和中国全国政协主席李瑞环亲切会见"一句，严格地说是不合语法的。因为"会见"是一个比较强的及物动词，表示施事接见受事，这受事一般不能表达成与事、并用介词引导而作状语。详细的讨论，请看前面《论元结构和句式结构互动的动因、机制和条件》一文的 §1.1。不过，这倒从反面向我们提出一个问题：在真实文本中，可能会出现一些不太符合一般的语法常规的表达。我们在做动词的论元结构研究，尤其是对动词及其支配的论元成分的配位方式进行描写时，怎样照顾到这种情况，以便计算机对真实文本作出语义关系处理。

关系，可以通过给从属成分标定论旨角色（thematic role）的方法来体现这种语义关系。因为真实文本中有关词项之间的关系主要通过谓词和其从属成分之间的语义关系来体现的，谓词及其论元之间的论元结构关系这种低层次的语义关系最终可以通过一定的程序传递到高层次的关于事件的脚本和框架结构中。其中，必须着重研究并突破两个难题：(i) 论元结构这种低层次的信息跟事件脚本或框架这种高层次的信息之间的传递和继承关系。比如，针对简历、生平介绍一类文本的信息抽取模板，必须设立出生、学习、工作、结婚、去世等子模板；于是在相应文本中诸如“出生、逝世、就学、毕业、工作、结婚”一类动词的论元都是比较重要的信息，是信息抽取时应该优先考虑的对象，并最终将成为填入模板中的信息项目。例如：

毛彦文简历

1898 年阴历 11 月 1 日出生于浙江省江山县城毛氏大家。辛亥革命后，她先后就读于江山西河女校……南京金陵女子大学。

1929 年秋，赴美国密歇根大学留学，主修中等教学行政。1931 年夏获硕士学位后……回国。回国后先后执教于复旦大学、暨南大学。

1935 年 2 月 9 日与前国务总理熊希龄结婚，并主持熊氏创办的北京香山慈幼院。…………

1949 年 4 月到台湾，1950 年赴美国。先后就职于旧金山“少年中国报”社、加州大学、华盛顿大学。

1962 年回台湾定居，并执教于实践家政专科学校，1966 年退休。现居于台北内湖。

（《中华读书报》，2000 年 10 月 11 日，第 5 版）

像上文中动词“出生”的主事论元“毛彦文”、时间论元“1898 年阴历 11 月 1 日”、处所论元“浙江省江山县城毛氏大家”、“就读”的处所论元“江山西河女校……南京金陵女子大学”等都可能是重要的、应该被抽取出来的信息项目。(ii)所指相同的名词性成分的论旨角色的保持或转变对于确定代词或空语类的所指的影响和作用。因为所指

相同的名词性成分对于同一个句子中、前后相邻的句子中的不同动词而言，其论旨角色可能是相同的、也可能是不同的；这种论旨角色的保持或转变对于句子中代词或空语类的所指和照应关系是有预测作用的。[①] 因此对于信息抽取来说，这种信息是比较关键和重要的。例如：

(5) 澳门报刊$_i$ 近来纷纷发表评论，〔e_i〕谴责美国对台出售先进武器，〔e_i〕奉劝美国政府在对台军售问题上悬崖勒马。(《每日电讯》，2001 年 4 月 27 日）

(6) 秘鲁《秘华商报》$_i$ 25 日发表评论，〔e_i〕严厉批评美国政府$_j$〔e_j〕不顾中国政府的强烈反对和严正交涉，〔e_j〕公然向台湾出售先进武器。（同上）

(7) 中国国家主席江泽民$_i$ 今天致电俄罗斯总统普京$_j$，〔e_i〕〔向 e_j〕祝贺俄罗斯国庆节。（《人民日报》，2001 年 6 月 13 日，第 1 版）

(8) 日前，云南省委宣传部、省科技厅$_i$ 组织 10 名优秀中青年学术带头人$_j$ 召开了“〔e_{i+j}〕纪念建党 80 周年，〔e_{i+j}〕共话科教兴国”的座谈会。（同上）

(9) 巴西联邦警察二十五日上午将在哥伦比亚被捕的巴西大毒枭费尔迪尼奥押解回巴西利亚，并暂时关押在联邦警察局的牢房里。（同 5）

在例(5)中，两个后续小句的空主语承先行小句的主语而省略，论元角色保持不变，都是施事。在例(6)中，第二个小句的空主语承先行小句的主语而省略，论元角色保持不变，都是施事；第二个小句中后续动词“不顾”的施事主语和第三个小句的主语跟先行动词“批评”的受事宾语同指，被强制性地删除了。例(7)第二个小句除了承先行小句而省略施事主语之外，还承先行小句的与事宾语而省略了状语位置上的与事及其介词。例(8)总体上是连谓结构作谓语的主谓结构，

① 这一点是董振东先生在“《信息处理用现代汉语词汇研究》课题组研讨会”(1998 年 12 月 22—24 日，北京)期间提醒我注意的，谨此致以诚挚的谢意。

在连谓结构的后段的动词“召开”的宾语中，有一个由两个述宾结构组成的并列结构；其中的动词的协同性施事正是前段动词“组织”的施事主语和受事宾语的复合体，即“云南省委宣传部、省科技厅”和“10名优秀中青年学术带头人”；并且，由于结构的原因，这种协同性施事只能以空主语的形式出现。例(9)则说明要快速而有效地抽取信息，还应该研究“押解回、关押在”等“V-回、V-在”类动词性结构的论元结构。

2.3 微观层次的逻辑结构知识(logic structure knowledge)

这基本上也是以句子为考察单位的。一般地说，句子的逻辑结构涉及到否定性词语(negative word)、量化词语(quantifier)、模态词语(modal word)，以及时体(tense, aspect)成分跟其所约束的成分之间的语义关系。下面，我们以否定关系为例进行讨论。

根据我们的考察，否定性词语对于确定文本中的事件到底发生与否和是非评价有决定性的影响，特别是否定词的辖域(scope of negation)到底管到哪儿、落入否定辖域中的哪些成分有可能成为否定词的否定焦点(focus of negation)等，这些因素对于信息抽取也是具有决定性的影响的。例如：

(10) 德国重申不参与售台(湾)武器。

(11) 我不在餐桌上批评孩子，以免大家消化不良。(《中华读书报》2001年3月14日，第4版)

(12) 他虽未声称，如果没有IBM大屠杀就不会发生，但指出受害者数目却由于IBM当时最好的技术而大为增加。(《中华读书报》2001年3月28日，第13版)
〔背景：布莱克在《IBM和大屠杀》(IBM and the Holocaust)中写道，纳粹迫害犹太人的各方面，无论是在人口普查中鉴定犹太人，还是在欧洲沦陷区管理苦役集中营，其速度提高都得益于IBM穿孔卡片分类器的运用。(出处同上)〕

像例(10)(11)中的“不”、例(12)中的“未”是信息抽取时不能丢掉的，否则将得到跟原文相反的意义。在例(11)中，虽然“在餐桌上批评孩

子”都落入“不”的否定辖域之中，但是只有作状语的焦点成分“在餐桌上”是真正被否定掉的，而中心语“批评孩子”的意义则保持下来了。也就是说，例(12)的真正意思是：{我批评孩子，但不在餐桌上做这种事}。例(10)中假设性的双重否定句“如果没有IBM大屠杀就不会发生”，基本上等同于“正是有了IBM，大屠杀才会发生的”。要正确地做到这种同义互释(paraphrase)，首先必须研究“如果……就……”格式可以表示假设和反事实(counterfactual)的特点，研究“没有……不……”等双重否定格式的语义表达功能。

汉语中有大量的由一连串小句(clause)组成的流水句(paratactic sentences)，出现在先行小句中的否定词到底管辖不管辖后续小句有时是两可的。例如：

(13) 这种类型的视感觉不像三色说所讲的，是由于不同颜色混合的结果。

(《感觉世界》中译本，第69页)

(14) 在我们看来，文革并不像林毓生、陈来教授说的那样，〔是“五四”反传统思想的继续和发展〕，恰恰相反，文革是“五四”对立面成分的回潮，……

(《北大中文研究》创刊号，第17—18页)

(15) 吕先生和许多严肃的学者一样，不喜欢随便上别人家去串门，〔把宝贵的时间虚掷在无谓的清谈之中。〕

(《中国语文》1998年第3期，第167页)

(16) 吕先生和许多严肃的学者一样，不喜欢随便上别人家去串门，〔把宝贵的时间都用在读书和做学问上。〕

单纯从结构上看，例(13)—(16)的后续小句既可以解释为在先行小句的否定词“不”的辖域之中，又可以解释为在“不”的辖域之外。从汉语的行文习惯上看，“……不像……”一类比况句的后续小句通常是落在“不”的辖域之中的，例(14)“恰恰相反”后的一句话证明了这一点；至于其他句式就只能全凭上下文来消除否定词的辖域歧义了，

比如例(15)和(16)的情况正好相反。[①] 这种情况将给计算机从后续句中抽取信息带来困扰。

一般地说,有无否定词会造成句子在语义表达上的肯定和否定的对立。但是,在某些句法环境之下,否定词似乎是冗余的(redundant),即否定词在语义功能上被某些特定的语法环境中和化(neutralization)了。例如:

(17) 在没有结婚之前,他对我是非常体贴的。

(18) 桐桐和她妈妈在没有下雨之前已经回去了。(笔者亲闻)

对于"在没有……之前"这类否定句,在信息抽取时几乎可以不理会其中的否定词"没有"。但是,在有些句式中,否定词到底影响不影响句子意义的肯定或否定表达,则取决于较为复杂的语用因素。例如:

(19) a. 曼联队差点儿赢了这场球赛(意为:没赢)

b. 曼联队差点儿没赢了这场球赛(意为:赢了/?没赢)

(20) a. 曼联队差点儿输了这场球赛(意为:没输)

b. 曼联队差点儿没输了这场球赛(意为:输了/没输)

从语义功能上看,"差点儿"相当于一个否定词,所以,用在肯定句中,整个句子的意思是否定的,如例(19)和(20)的 a 句所示;用在否定句中,整个句子的意思是肯定的,如例(19)和(20)的 b 句所示。同时,例(19)和(20)的 b 句还有一种否定的释义,即"差点儿+没+VP"是一种歧义格式;至于它到底表示肯定意义还是否定意义,在很大程度上取决于说话人的期望:当他希望 VP 所表示的事件发生时,"差点儿+没+VP"表示肯定意义;当他不希望 VP 所表示的事件发生时,"差点儿+没+VP"表示否定意义,这时否定词"没"基本上是冗余的。[②] 这种否定句的复杂情形,给信息抽取带来了莫大的困扰。

我们相信,一个较大规模的、标注了上述三种语义关系的语料库

① 详细的情况,请看袁毓林(2000)。

② 详细的情况,请看朱德熙(1980)§3.2。

(即网库),对于信息抽取工作是十分有用的,不管它是用基于(语言学)规则的方法还是用基于统计的方法。

3 下一步工作的方案和思路

上文讨论了哪些语义知识对信息抽取是重要的,现在我们简要地说明下一步怎样来研究这些语义知识,以及怎样来实施这些工作。

3.1 工作方案

由于这种课题主要研究怎样为信息抽取提供语义知识方面的资源,因而所选的汉语文本最好已经有了初步的语法标注(包括分词、词类标注、甚至短语边界等信息)。为此,我们打算从北京大学计算语言学研究所加工出来的 1998 年《人民日报》语料(已经完成了词语切分、词类标注和专名处理)中,选择诸如“职务调动、产品发布、个人简历、自然灾害、新型材料、企业重组、投资方向、消费结构、旅游经济、假日消费、公费医疗、休闲方式”等几十个专题的文本,进行篇章关系、题元角色、角色转换、照应关系、否定关系等语义信息的标注,并研究这些标注内容怎样有效地为信息抽取提供语义知识方面的支持。

3.2 研究思路和方法

由于这种课题主要研究怎样通过对真实的汉语文本进行语义标注,来为计算机自动地从汉语文本中抽取出指定信息提供语义知识方面的资源;因而在对语义知识的揭示、表达和组织等各个环节上,都需要有合理并且可以形式化的语义学理论和方法作支撑。为此,我们打算首先采用话语分析(discourse analyses)和篇章语言学(text linguistics)的理论和方法,结合传统语法中的复句研究和汉语句群研究的成果;对同一个句子中不同的小句、同一句群中的不同句子之间的语义关系进行分析,着重找出相应的语法形式标志。用确定中心理论(Centering Theory)来研究语流中不同句子的注意焦点(focus of attention)的变化(包括继续、保存和转变)及其追踪机制

(tracking mechanism)和不同的话语片断中小话题(discourse segment topic)的转变机制。接着用依存语法(dependent grammar)和配价语法(valence grammar)的思想来发现谓词性成分(包括有价名词)及其从属成分之间的依存关系,并确定不同的谓词性成分的价数;用格语法(case grammar)和论元结构理论(argument structure theory)来确定这些从属成分的论旨角色及其句法配置方式;并用框架语义学(frame semantics)中"框架—槽"的形式表示出来。然后用档案卡更新语义学(file change semantics)和篇章表示理论(discourse representation theory)来分析所指相同的名词性成分对于不同的谓词其论旨角色的保持或转变,确定无定名词短语(indefinite NP)的语义所指、代词和空语类跟先行词的照应关系及其语义解释。最后用形式语义学(formal semantics)关于否定句的三分结构(tripartite structure)的理论来刻画否定句的语义结构,分析否定的辖域和焦点对于信息抽取的影响;用数理逻辑(mathematical logic)和生成语法(generative grammar)中的量化理论(quantification theory)来分析带量化词的结构,用模态逻辑(modal logic)和时间逻辑(time logic)来分析句子的模态和时体。

参考文献

顾　阳 (1994)《论元结构介绍》,《国外语言学》第1期。

蒋　严、潘海华 (1998)《形式语义学引论》,中国社会科学出版社。

孙　斌 (2000)《继承—归纳机制及其在对象系统中和信息提取技术中的应用》,北京大学计算机系博士学位论文。

徐烈炯 (1990)《语义学》,修订本1995年,语文出版社。

袁毓林 (1998)《汉语动词的配价研究》,江西教育出版社。

袁毓林 (2000)《流水句中否定的辖域及其警示标志》,《世界汉语教学》第3期。

朱德熙 (1980)《汉语句法中的歧义现象》,《中国语文》第2期。

Bosch, Peter & Rob van der Sandt (1999) *Focus: Linguistic, Cognitive, and Computational Perspective*, Cambridge University Press.

Jackendoff, Ray (1990) *Semantic Structure*. MIT Press.

Lappin, Shalom (1997) *The Handbook of Contemporary Semantic Theory*,

Blackwell Publishers.

Leech, Geoffrey (1981) *Semantics: The Study of Meaning*. Penguin Books. 《语义学》,李瑞华等译,上海外语教育出版社,1987年。

Lyons, John (1977) *Semantics*, Vol. 1 & 2. Cambridge University Press.

Pan Haihua (2001a) Focus and Scope Interaction in the Interpretation of *Bu*-Sentences in Mandarin Chinese, Lectures in Peking University, June 5, 2001.

Pan Haihua (2001b) Centering Theory and Focus Tracking in Discourse, Lectures in Peking University, June 12, 2001.

Steinberg, Danny & Jakobvits, Leon (1976) *Semantics: An Interdisciplinary Reader in Philosophy, Linguistics and Psychology*. Cambridge University Press.

2001年5月初稿,2002年6月改定

(发表于《中文信息学报》2002年第5期)

用动词的论元结构跟事件模板相匹配

——一种由动词驱动的信息抽取方法

本文以孙斌(2000)信息抽取模型(InfoX)的测试语料(职务变动文本)为主要对象,具体说明怎样建立从动词的论元结构到相关的事件模板的匹配关系。首先根据职务变更动词的有关句法、语义特点,把它分成六个小类:任命、担任、免职、辞职、调遣、受命;然后,分别描写每一小类动词的论元结构,特别是它们所支配的论元角色及其句法配置方式。最后,建立动词的论元角色跟事件模板元素的匹配关系,并揭示动词对文本筛选和合并都有导向作用,说明发展由动词驱动的信息抽取方法的可行性。

1 信息抽取模板和动词的论元结构

在信息抽取(information extraction,简称 IE)系统中,信息抽取模板起到把要提取的信息内容类型化和结构化的作用。比如,用户所关心的一个任职事件中的四个信息项目:谁、什么时候、什么组织、什么职务,可以表示为任职模板中的四个模板元素(template elements)。这样,跟某种特定事件相关的模板就是一个事件模板,模板中的槽(slots)就是事件的参与者(participants)。如果把一个事件模板看作是一个句子的语义的某种抽象化表示,那么模板元素之间的关系就是动词的及物性关系意义(transitivity),各个模板元素就是动词所支配的论元(argument)。因此,袁毓林(2002)指出:动词的论元结构可以传递到事件脚本或框架中,动词的论元最终将成为填入事件模板中的信息项目(第 10 页)。

本文以孙斌(2000)应用实例"职务变动(succession)"的测试语料为主要对象,具体说明怎样建立从动词的论元结构到相关的事件模板的匹配关系。

2 职务变动文本中动词的类型和特征

孙斌(2000)选取北京大学计算语言学研究所加工出来的1998年《人民日报》语料，对他设计的信息抽取模型(InfoX)进行测试。这是一种已经完成了分词、词类标注和专名处理的语料。他通过人工阅读头两个月的语料(约17.5MB文本)，从中找出70多个职务变动事件，其中“任职”45个，“离职”16个，“调职”11个。召回率(正确数目除以实际数目)为45%，基本达到其设计目标。

我们对这些测试语料(47个文本)重新进行了分析，发现其中实际出现了81个职务变动事件，每一个事件都由一个动词及其从属成分来表达。为了方便，这种表示职务变动的动词可以叫作“职务变更动词”，相应地，表示任职的动词是“任职动词”，表示离职的动词是“离职动词”，表示调职的动词是“调职动词”。下面是这种动词的类型(type)及其实例(token)，括号中的数字表示使用的次数：

(1) 任职动词，共19+2个，分为“任命”和“担任”两类：

A. 任命动词：任命(10)、选(2)、选举(2)、提名(1)、选聘(1)

B. 担任动词：任(18)、担任(10)、就任(1)、出任(2)、上任(1)、现任(1)、连任(2)、历任(1)、兼任(1)、就职(3)、当(2)、当选〔为〕(6)，接任、继任

(2) 离职动词，共2+8个，分为“免职”和“辞职”两类：

A. 免职动词：免去、撤销、撤消、撤除、解除、罢免、免职、撤职

B. 辞职动词：辞去(3)、辞职(2)，离任、下台、任满

(3) 调职动词，共3+1个，分为“调遣”和“受命”两类：

A. 调遣动词：调〔动〕……任……、升〔为〕(1)、提升〔为〕(1)

B. 受命动词：调任(3)

可见，描述职务变动的动词之数量是有限的，孙斌(2000)只涉及到24个。我们在网上搜索了大量的职务变动文本，补充了11个(后面没有使用次数的)。这样，总共才35个。更为重要的是，这些动词不仅可以分成任职、离职和调职三类；而且，每一类下面都可以分成对称的使动和自动两类：任命～担任、免职～辞职、调遣～受命。于

是，这六类动词便只有两套句法上的组配模式（详见下一节），为信息抽取提供了简明扼要的语言表达模式。另外，这三大类、六小类动词又是由有限的几个语素组成的。比如，任命动词主要用“选”、担任动词主要用“任”、免职动词主要用“免、撤”、辞职动词主要用“辞”、调职动词主要用“调”。并且，这些语素都是独立用一个汉字来书写的，辨识起来十分容易。这些因素都非常有利于实行由动词驱动的信息抽取的路线。

最后，我们发现孙斌（2000）的文本中还用了比较抽象的“成〔了〕（1）、成为（3）、是（4）”来表达担任行为，用终止动词“退休（1）、病逝（1）”来隐含离职行为。这种非专用的职务变更动词只占 10/81。也就是说，专用的职务变更动词占 70/81（80％弱）。对于非专用的职务变更动词，需要在篇章中依赖上下文语境来帮助推导。这些问题，我们将有另文讨论。

3　职务变更动词的论元结构

一般来说，动词的论元结构包括动词的论元属性（可以带多少论元，即配价数目）、论旨属性（这些论元跟动词的语义关系，即语义格）、范畴属性（这些论元分别由什么词类来实现）、句法属性（这些论元分别充当什么句法成分）、语义属性（实现这些论元的词语有什么语义特征）、句法配置（动词跟其论元可以构成哪些句法格式）。对此，下面采用诸动词共有的合叙，殊异的分述这种权宜的、节省篇幅的体例。

职务变更动词一般带有系事（relative，记作 Re）和经事（experiencer，记作 Ex）两种论元角色，分别表示某种职务和担任或辞去这种职务的人；有的还带有施事（agent，记作 A），指实施任命或解除这种职务的人或组织。① 例如：

（1）［周小川］$_{Ex}$出任［中国人民银行行长］$_{Re}$。（南方网）

① 为了方便，动词的论元的论旨角色简称论元角色。关于论元角色的类型及其句法、语义特征，详见袁毓林（2002b，2003）。

(2) [张云川]$_{Ex}$辞去[湖南省长职务]$_{Re}$。(南方网)

(3) [中央军委]$_{A}$任命 [孔英]$_{Ex}$为 [陕西省军区政委]$_{Re}$。(南方网)

(4) [中共]$_{A}$免去 [张文康]$_{Ex}$[卫生部党组书记职务]$_{Re}$。(南方网)

其中,施事 A 由名词性成分充当,其语义特征(约束条件)是指人或组织(记作 NP[+human / organization]),其句法属性是一般作主语;经事 Ex 由名词性成分充当,其语义特征是指人(记作 NP[+human]),其句法属性是一般作宾语(在施事出现的情况下),或主语(在施事不出现的情况下);系事 Re 由名词性成分充当,其语义特征是指职务(记作 NP[+position]),其句法属性是一般作宾语。

下面,分别描写三大类六小类职务变更动词的论旨结构及其句法配置方式。

3.1 任命动词的论旨结构

任命动词的论旨结构可以表示为:V[+appoint]:{A, Ex, Re}。其句法配置主要有两种格式:

S_{1a}: A+__+Ex+为/当/任/担任+Re

S_{1b}: Ex+被[A]+__+为/当/任/担任+Re

方括号表示其中的成分有时可以省略。介词"为"作为格标记(case marker),应该跟其所标记的系事格连在一起;但是它可以跟"当、任、担任"等动词交替使用,所以变通处理为动词。S_{1b}是S_{1a}的被动转换格式,所有的任命动词都可以进入这两种格式。例如:

(5) [墨西哥总统塞迪略]$_{A}$ 任命[女参议员罗萨里奥·格林]$_{Ex}$为[外交部长]$_{Re}$。(孙,35)[①]

(6) [徐虎]$_{Ex}$被任命为[这所学校的校长]$_{Re}$。(孙,3)

(7) [大连市第十二届人民代表大会第一次会议]$_{A}${1 月 10

① 例句后括号中的"孙"表示引自孙斌(2000)的测试语料,数字代表语料的文本编号;这编号原文没有,是我们为了查找和核对的方便而加上去的。

日}$_T$ 选举[于学祥]$_{Ex}$为[市人大常委会主任]$_{Re}$，选举[薄熙来]$_{Ex}$为[大连市市长]$_{Re}$。(孙,39)

(8) [木庭健太郎]$_{Ex}$被选为[干事长]$_{Re}$。(孙,9)

(9) 塔拉尔$_i$ 于1997年3月当选巴参议院议员，[e_i]$_{Ex}${同年12月15日}$_T$ 被[执政党穆斯林联盟谢派]$_A$ 提名为[总统候选人]$_{Re}$。(孙,11)

其中，T代表时间(time)这种动词的非核心论元。e代表空语类(empty category)，即被省略掉的成分，下标表示这个空语类跟前面有相同下标的名词性成分所指相同。

3.2 担任动词的论旨结构

担任动词的论旨结构可以表示为：V[+hold]：{Ex, Re}。其句法配置主要有四种格式：

S_{2a}：Ex+__+[为]Re

S_{2b}：由Ex+__+Re

S_{2c}：Ex+__

S_{2d}：Re+由Ex+__

S_{2c}是S_{2a}的省略形式，S_{2d}是S_{2b}的变换形式。所有的担任动词都可以进入S_{2a}，除了"当选"可以后加介词"为"，其他都不行。只有"上任、当选、就任、就职、连任、接任、继任、获任"等双音节动词可以进入S_{2c}。只有"担任、出任、兼任、接任、继任"等少数双音节动词可以进入S_{2d}。例如：

(10) {1996年初}$_T$，[李长水]$_{Ex}$担任了[市公安局长、党委书记]$_{Re}$。(孙,5)

(11) [刘沈明]$_{Ex}$原{在福建省海洋渔业公司}$_L$ 当[车间主任]$_{Re}$。(孙,4)

(12) 香港特区政府昨天公布了香港特区基本法推广督导委员会成员名单，[政务司长陈方安生]$_{Ex}$出任[委员会主席]$_{Re}$，[高苕华]$_{Ex}$任[副主席]$_{Re}$。(孙,2)

(13) 邢云$_i$,1952年生,大学文化。[e_i]$_{Ex}$历任[内蒙古伊克昭盟副盟长、盟委副书记、盟长]$_{Re}$,{1996年10月起}$_T$任[盟委书记、盟人大工委主任]$_{Re}$。(孙,23)

(14) 1938年2月,中共晋冀豫省委在太岳区沁县设立办事处,由[省委统战部部长安子文]$_{Ex}$兼任[办事处主任]$_{Re}$。(孙,28)

(15) [原中联办主任姜恩柱]$_{Ex}$获任[人大外事委副主任委员]$_{Re}$。(南方网)

(16) [现任信息产业部部长王旭东]$_{Ex}$接任[国信办主任一职]$_{Re}$。(南方网)

(17) 摩洛哥新一届两院制议会7日选出第一任参议院议长,[原摩洛哥一院制议会议长艾赛义德]$_{Ex}$当选为[第一任议长]$_{Re}$。代表院议长$_i$已于6日选举产生,[原议会第一副议长拉迪]$_{Ex}$当选[e_i]$_{Re}$。(孙,34)

(18) [巴基斯坦穆斯林联盟谢里夫派候选人、原最高法院大法官穆罕默德·拉斐克·塔拉尔]$_{Ex}$,{今天}$_T${在巴国民议会、参议院以及各省议会选举中}$_L$,当选[巴基斯坦第九任总统]$_{Re}$,任期5年。(孙,1)

其中,L代表处所(location)这种动词的非核心论元。

3.3 免职动词的论旨结构

免职动词的论旨结构可以表示为:V[+remove]:{A, Ex, Re}。其句法配置主要有四种格式:

S_{3a}:A+__+Ex[的]+Re[的]职务

S_{3b}:Ex+被[A]+__+Re[的]职务

S_{3c}:A+__+Ex的职务

S_{3d}:A+对Ex+__

S_{3b}是S_{3a}的被动转换格式,S_{3c}是S_{3a}的省略形式。"免去、撤销、撤消、撤除、解除、罢免"等及物动词都可以进入S_{3a-c},不能进入S_{3d};"免职、撤职"等不及物动词只能进入S_{3d},不能进入S_{3a-c}。例如:

(19) 新华社在会后不久就宣布，[中共中央]$_A$ 已经撤除[张文康]$_{Ex}$[在卫生部的党职]$_{Re}$。(雅虎中国)

(20) [咸阳市人大常委会]$_A$ 决定撤消[张定会]$_{Ex}$[副市长职务]$_{Re}$。(中新网)

(21) 鄢良钟$_i$ 原任四川省内江市市长，[他$_i$]$_{Ex}${因接受贿赂}$_{Rn}$被依法撤消职务。四川省十届人大常委会四次会议$_j$ 于7月25日通过罢免案，[e$_j$]$_A$ 罢免[鄢良钟]$_{Ex}$[十届全国人大代表职务]$_{Re}$。(新华网)

(22) 俄罗斯总统普金$_i$23日签署命令，[e$_i$]$_A$ 解除[亚历山大·阿夫杰耶夫]$_{Ex}$的[第一副外长职务]$_{Re}$，同时任命瓦列里·洛希宁为第一副外长。(新华网)

(23) 9月7日，铁道通信信息有限责任公司发生重大人事变动。[原总经理彭朋]$_{Ex}${经董事会召开临时会议}$_M$，被解除职务，新任铁通公司总经理由乔金洲担任。(新浪网)

(24) [田凤山]$_{Ex}${因违纪}$_{Rn}$被免去[国土资源部部长职务]$_{Re}$。(中国网)

其中，Rn和M分别代表原因(reason)和方式(manner)这两种动词的非核心论元。

3.4 辞职动词的论旨结构

辞职动词的论旨结构可以表示为：V[+resign]：{Ex, Re}。其句法配置主要有四种格式：

S_{4a}：Ex+__+Re[的]职务

S_{4b}：Ex+__+Re

S_{4c}：Ex+__+职务

S_{4d}：Ex+__

S_{4b-c}是S_{4a}的省略形式。“辞去”等及物动词都可以进入S_{4a-c}，不能进入S_{4d}；“辞职、离任、下台、任满”等不及物动词只能进入S_{4d}，不能进入S_{4a-c}。例如：

(25) 墨西哥恰帕斯州议会7日批准[胡里奥·鲁依斯]$_{Ex}$辞去[州长职务]$_{Re}$，并任命众议员罗伯特·阿尔沃雷斯·纪廉为新州长。(孙,40)

(26) [以克劳斯为首的捷克原政府]$_{Ex}$是{一九九七年十一月三十日}$_{T}$被迫辞职的。(孙,10)

(27) 驻港部队司令员调整，[熊自仁]$_{Ex}$离任，王继堂接任。(南方网)

3.5 调遣动词的论旨结构

调遣动词的论旨结构可以表示为：V[＋dispatch]：{A，Ex，Re}。其句法配置主要有五种格式：

S_{5a}：A＋__＋Ex＋当/任/担任＋Re

S_{5b}：A＋把 Ex＋由 Re_1＋__＋为 Re_2

S_{5c}：Ex＋被 A ＋由 Re_1＋__＋为 Re_2

S_{5d}：Ex[的职务]＋__＋为 Re

S_{5e}：Ex[的职务]＋由 Re1＋__＋为 Re_2

其中，Re_1 表示原来的职务，Re_2 表示变更后的职务。S_{5d-e}可以看作是 S_{5c}的省略形式。只有“调”等少数动词可以进入 S_{5a}，只有“升、提升、提拔”等动词可以进入 S_{5b}和 S_{5c}。在实际的新闻语料中，S_{5a-c}这些格式的用例并不多见，常见的是 S_{5d-e}这些格式的用例。例如：

(28) {根据越南国家主席和政府总理的决定}$_{M}$，[国防部长范文茶]$_{Ex}$由[中将]$_{Re1}$提升为[上将]$_{Re2}$。(孙,14)

(29) [邱娥国]$_{Ex}$的职务虽已升为[分管户籍、外勤的副所长]$_{Re}$。(孙,45)

3.6 受命动词的论旨结构

受命动词的论旨结构可以表示为：V[＋transfer]：{Ex，Re}。其句法配置主要有两种格式：

S_{6a}：Ex＋由/从 Re_1＋__＋为 Re_2

S_{6b}：Ex＋__＋Re

只有“调任”等少数动词可以进入 S_6。例如：

(30)［诸葛彩华］$_{Ex}$ 从［县委副书记岗位］$_{Re1}$ 调任［代县长］$_{Re2}$。(孙,42)

(31)［史有彪］$_{Ex}$｛1987 年 8 月｝$_T$ 调任［市委农工部副部长、农业委员会副主任］$_{Re}$。(孙,44)

4 论元角色和模板元素的对应关系

孙斌(2000)的职务变动信息抽取模型中共有四个特征(必备格)：时间(Time)、组织(Org)、职务(Post)和人物(Person)。其中，人物在任职事件(Start_job)中具体化为任职者(Who_in),在离职事件(Leave_job)中具体化为离职者(Who_out),在调职事件(Change_job)中则既是前一个职务的离职者,又是后一个职务的任职者。这四个事件模板元素,基本上都能在相应的职务变更动词的论元结构中找到。

一般地说,人物就是经事(Person ← Ex),少数经事在作后续句的主语时可以承上文而省略(如：9、13)。职务就是系事(Post ← Re),调职事件的前一个职务是系事 Re1、后一个职务是系事 Re2。少数系事可以承上文而省略(如：17、21、23)；至于“辞职、离任”等极少数辞职动词,其系事一定不能在本小句中出现(如：26、27)。时间就是时间论元(Time ← T)。时间 T 属于动词的非必有(optional)论元,一般不必在论元结构中表示出来。因为,一方面它基本上是对所有的动词都开放的,即每一个动词都可以拥有一个时间论元；另一方面它的句法位置是比较固定的、可以预测的,通常在动词之前,并且要么在第一个论元之前(如：例 10)、要么在第一个论元之后(如：例 7、9、13、18、26、31)。当然,也可以承上文而省略(如：例 14、21—23、25)。组织一般在系事论元中充当职务名词的修饰语(如：例 5—7、10、12—16、18、19、21、24、31)。也可以作为动词的处所这种非必有论元,在动词前面以状语的句法身份出现(如：11)。当然,也可以

承上文而省略(如：8、9、17、20、23、25、27、30)。

为了概括，可以根据所能支配的必有论元的数目和类别，把上文六类职务变更动词分为两大类：(i) 使动类，包括任命、免职、调遣三类动词。其特点是：有{A, Ex, Re}三个论元角色，其主要的句法格式是 S_i：[T+] A+__+Ex+[当/任+] Re，相应的事件框架为 E_i：[Time+] A+__+Person+[当/任+] Org-Post。(ii) 自动类，包括担任、辞职、受命三类动词。其特点是：只有{Ex, Re}两个论元角色，其主要的句法格式是 S_{ii}：[T+] Ex+ __+Re，相应的事件框架为 E_{ii}：[Time+] Person+__+Org-Post。当然，为了准确，在信息抽取系统的分析词典中，每一个职务变更动词的论元结构都应该得到充分的描写。

孙斌(2000)的信息抽取模型(InfoX)的系统结构中，有一个"Names 识别"模块；只要找出人名，就可以用它来匹配事件模板中的人物这一模板元素。我们设想，对于处理职务变动文本的系统，还应该设置三个识别模块：一个"Post 识别"模块；只要找出职务名词，就可以用它来匹配事件模板中的职务这一模板元素。并且，基本的职务名词是有限的、可列举的；派生的职务名词又是有规则的，如"一长、副一、代一"等。一个是"Org 识别"模块；只要找出组织、机构名词，就可以用它来匹配事件模板中的组织这一模板元素。并且，常用的基础的组织、机构名词是有限的、可列举的；派生的组织、机构名词又是有规则的，如"一党、一国/省/州/市/县、一部/厅/局/处、一院/所/室、一厂/公司、一委员会/理事会/董事会"等。有了对人名和职务、组织名词的正确识别，那么包含"是、成为"等抽象动词的句子也容易处理了。其约束条件是：当其前面的成分是人名、后面的成分是职务名词时，这些成分的论元角色就分别是经事和系事。例如：

(32) [龚德俊]$_{Ex}$ 是[北京中诚信租赁有限公司的董事长]$_{Re}$。(孙,16)

(33) 1992 年，刘涛$_i$ 进入江西农用机械厂，[e_i]$_{Ex}$ 成为[总工程师]$_{Re}$。(孙,47)

最后一个是“Time识别”模块，只要找出时间名词，就可以用它来匹配事件模板中的时间这一模板元素。并且，基本的时间名词的构造形式是有规则的；比如，“—年—月—日”。指代性的时间名词又是可列举的，比如“今天、去年、日前、此后”等。

5 结语：动词的引导作用贯穿全过程

信息抽取系统一般都有一个预处理过程，以过滤掉文本中跟抽取目标无关的句子（达90%左右），然后通过词法分析来识别跟抽取目标有关的词汇，即“关键词”识别和标引；再对包含关键词的句子作句法、语义分析，找出相关的数据填入数据模板。[①] 拿职务变动文本来说，其关键词显然是职务变更动词。那么，有了上文讨论的职务变更动词的分类及其语义特征、每一小类动词所支配的论元角色及其语义约束和句法配置方式这些论元结构知识，再加上动词的论元角色跟事件模板元素的匹配关系的知识；势必会提高信息抽取系统工作的精确性，包括召回率（正确数除实际数）和正确率（正确数除抽取数）。这样，对含有职务变更动词的句子作浅层分析的主要目标便是：找出相关动词及其支配的论元角色，确定有关短语的边界，把句子中有关的命名实体（named entities）跟句子中动词的论元角色对应起来，为把论元角色跟模板元素匹配作准备。

当模板匹配完成以后，就进入后处理阶段：对每个匹配出来的实例作进一步的检查和修正，补足相应模板中空缺的槽；再调用一个综合处理函数把有关实例合并起来，形成具体的表示。[②] 通过调查，我们发现这种实例合并工作仍然可以用动词来驱动。原则是：如果几个句子用了同一个或同一小类的动词，并且其论元角色（包括时间等非必有论元）是同指的；那么，它们就是同一个事件的不同的语句表达，应该合并起来。例如：

(34) [在肯尼亚大选中赢得连任的肯尼亚总统莫伊]$_{Ex}$ {5

① 详见孙斌(2000)，第106—107页。

② 详见孙斌(2000)，第116—117页。

日}$_T$ 在内罗毕宣誓就职。据肯尼亚选举委员会昨天正式宣布，在去年12月29日至30日举行的大选中，肯尼亚非洲民族联盟候选人、现任总统丹尼尔·阿拉普·莫伊$_i$以较大优势击败了14名反对党候选人，[e_i] Ex 再次当选为[肯尼亚总统]$_{Re}$，任期5年。（孙，20）

（35）陕西省咸阳市第四届人民代表大会常务委员会第26次会议$_i${13日}$_T$ 通过了{[关于（[e_i]$_A$ 撤消[张定会]$_{Ex}$[咸阳市人民政府副市长]$_{Re}$）]的}决定。……咸阳市人大常委会$_i$决定[[e_i]$_A$ 撤消[张定会]$_{Ex}$[副市长职务]$_{Re}$。（新华网）

例(34)中先后用了"连任、就职、当选"三个担任动词，但其论元角色的所指相同；因此，可以合并成一个表示。例(35)先后用了同一个动词，并且其论元角色的所指相同；因此，完全可以合并成一个表示。

可见，从开始预处理时的关键词识别，到中间的模板选择和模板元素匹配，到最后的后处理时把表达同一事件的实例合并表示，相关动词一直起着驱动和引导作用。因此，这种以动词为主导的信息抽取路子可以称为动词驱动的信息抽取方法。

当然，光靠上文涉及的动词语义所表示的事件类型及其论元结构知识，显然是不够的；还应该考虑基于论元结构的篇章和逻辑知识，来更准确地确定事件的信息类型以及有关信息特征的分布位置。这正是我们另一篇文章的主题。

参考文献

顾　阳（1994）《论元结构介绍》，《国外语言学》第1期。

孙　斌（2000）《继承—归纳机制及其在对象系统中和信息提取技术中的应用》，北京大学计算机系博士学位论文。

袁毓林（1998）《汉语动词的配价研究》，南昌：江西教育出版社。

袁毓林（2002a）《信息抽取的语义知识资源研究》，《中文信息学报》第5期。

袁毓林（2002b）《论元角色的层级关系和语义特征》，《世界汉语教学》第3期。

袁毓林（2003）《一套动词的论元角色的语法指标》，《世界汉语教学》第3期。

2004年5月初稿，2004年9月改定

（发表于《中文信息学报》2005年第5期）

用逻辑和篇章知识来约束模板匹配

——逻辑结构和篇章结构知识在信息抽取中的运用

本文以孙斌(2000)的语料为主要对象,讨论语句的逻辑结构和篇章结构怎样约束信息模板的类型,并约束对当前句中缺失的或以代词等形式表达的信息项目的求解。首先说明什么是基于论元结构的逻辑结构和篇章结构知识,然后分析否定算子、时体成分怎样改变事件的类型及其跟有关事件模板的匹配关系。接着,讨论动词的论元结构的内嵌和名词化等句法操作,怎样造成有关论元及相应的信息项目的分布位置发生变化。最后,讨论怎样利用篇章结构知识来求解本句中缺失的或以代词、指示词形式表达的信息项目。

1 基于论元结构的逻辑结构和篇章结构

袁毓林(2002)指出,除了论元结构知识之外,篇章结构和逻辑结构知识对信息抽取(information extraction)也有十分重要的作用。袁毓林(2005)具体讨论了怎样建立从动词的论元结构到相关的事件模板的匹配关系。本文打算进一步说明语句的逻辑结构怎样约束信息的类型及其跟有关的事件模板的匹配关系,篇章结构怎样约束当前句中缺失的或以代词、指示词等形式表达的信息项目及其跟有关模板元素的对应关系。

由于我们强调动词驱动的、以论元结构为基础的信息抽取路线,因而对语句的逻辑结构和篇章结构的分析势必也是以论元结构为基础的;比如,它们怎样帮助确定论元结构所反映的信息类型,找回当前论元结构中缺失的论元。这样,我们描写语句的逻辑结构和篇章结构就有了一个明确的目标和参照,逻辑结构便是附加在论元结构之上的否定、时体和模态等逻辑算子及其变量之间的语义约束关系,篇章结构便是从前一个论元结构到后一个论元结构的推进和关联,

特别是论元(包括非必有论元)的传递、称代和省略。为了方便,可以称这种分析逻辑结构和篇章结构的方法为基于论元结构的逻辑结构和篇章结构分析法。

本文继续以孙斌(2000)的应用实例"职务变动"的测试语料为主要对象,具体说明怎样用基于论元结构的逻辑结构和篇章结构知识来约束模板匹配、找全模板元素。

2 用逻辑结构知识约束信息的类型及其模板匹配

袁毓林(2005)指出,在职务变动文本中,不同类别的职务变动事件由不同类别的职务变更动词来表达。具体地说,任职动词表达任职事件,离职动词表达离职事件,调职动词表达调职事件。因此,动词的类别可以决定事件模板的类型。

事实上,动词的类别跟事件模板的类型之间的对应关系,常常会受到语句的逻辑结构的影响而发生扭曲;当然,这种扭曲关系也是有规律的。常见的情况有:

2.1 否定算子改变了事件的类型,使之适合于跟动词意义相反的事件模板。例如:

(1) 黄卫任建设部副部长,不再担任江苏省副省长职务。(南方网)

否定副词"不"使得"担任"表示的任职事件转变为离职事件。据此,可以得出规则:如果担任动词之前有否定算子,那么整个句子表示离职事件。

2.2 时体算子改变了事件的类型,使得句子找不到合适的事件模板。例如:

(2) 阿夫杰耶夫将出任俄罗斯驻法国大使。(新华网)

表示将来时的副词"将"使得"出任"表示的任职事件失去了现实性(irrealis),因而无法找到合适的事件模板。据此,可以得出规则:如果职务变更动词前有表示将来时的词语,那么没有合适的事件模板

可供匹配。

2.3 时体算子改变了事件的类型,使之适合于跟动词意义相反的事件模板。例如:

(3) 刘沈明原在福建省海洋渔业公司当车间主任,下岗5年多,他到处打工。(孙,4)

(4) 博塔曾任旧南非的国防部长、总理和总统。(孙,33)

(5) 他们都曾担任原公明党副书记。(孙,10)

(6) 塞维里诺是菲律宾的一位外交家,曾先后担任菲律宾驻美国、中国和马来西亚等国的外交使节。(孙,21)

(7) 列希曾连续担任过十一届议员,……(孙,15)

(8) 龚德俊是北京中诚信租赁有限公司的董事长,曾在中汽专用汽车珠海制造有限公司任过总经理。(孙,16)

时间副词"原、曾"和时体助词"过"表示过去有过某种行为或状况,含有现在已经不是这样的意思。[①] 因此,它们使得"当、任、担任"等动词表示的任职事件转变为离职事件。据此,可以得出规则:如果担任动词之前有"原(来)、曾(经)"等表示过去的时间副词或之后有表示经历体的助词"过",那么整个句子表示离职事件。

2.4 时体算子改变了事件的类型,使单一事件变成复合事件。例如:

(9) 原北大副校长陈章良任中国农大校长。(南方网)

(10) 原中联办主任姜恩柱获任人大外事委副主任委员。(南方网)

(11) 巴基斯坦穆斯林联盟谢里夫派候选人、原最高法院大法官穆罕默德·拉斐克·塔拉尔,今天在巴国民议会、参议院以及各省议会选举中,当选巴基斯坦第九任总统,任期5年。(孙,1)

(12) 摩洛哥新一届两院制议会7日选出第一任参议院议

---

① 详见吕叔湘主编(2001),第111—112、247、638—639页。

长，原摩洛哥一院制议会议长艾赛义德当选为第一任议长。代表院议长已于6日选举产生，原议会第一副议长拉迪当选。（孙，34）

担任动词的经事论元中有职务名词作同位性定语，并且这个定语自己还有区别词“原”作修饰语；同一个经事既有以前的（刚卸任的）职务、又有新任的职务，一个担任动词句表达了同一个人物前后相继地卸任原职务、担任新职务，即为调职。[①] 据此，可以得出规则：如果任职动词的经事的修饰语之前有区别词“原”，那么整个句子表示调职事件。

也就是说，经事论元中的“原”可触发调职模板。可资比较的是下面这些句子：

(13) 香港特区政府昨天公布了香港特区基本法推广督导委员会成员名单，政务司长陈方安生出任委员会主席，高苕华任副主席。（孙，2）

(14) 南非总统曼德拉已任命南非驻中国研究中心主任戴克瑞为首任驻华大使。（孙，12）

例(13)(14)跟(9)—(12)句式基本一样，只是缺一个“原”字；就有两种可能性：或者表示兼任，或者表示调职。但是，根据语言交际的缺省(default)规约，[②]可以把这种不用“原”的无标记形式看作是任职（兼任）事件。

3 论元结构的转换和信息项目的分布

袁毓林(2005)给出了职务变更动词的论元结构，特别是动词跟其论元角色的句法配置方式，希望为信息抽取系统提供有关的信息

① 孙斌(2000：121)指出，一个“调职”事件是具有某些共性（约束）的两个“任职”和“去职”事件的归纳，这两个至少需要满足的约束条件是：具有相同的 Person 值；Time 值前后相继。

② 关于缺省约定，详见袁毓林(1998)第26、114、139—142页。

项目的分布位置。必须指出,当论元结构经历了内嵌(embedding)、名词化(nominalization)等语法过程,从而处于从属的(subordinate)句法地位时;论元角色及其对应的有关信息项目的位置,也会相应发生有规律的变化。下面分项讨论。

3.1 首先要剔除跟动词同形的其他词类。例如:

(15) 据肯尼亚选举委员会昨天正式宣布,在去年12月29日至30日举行的大选中,肯尼亚非洲民族联盟候选人、现任总统丹尼尔·阿拉普·莫伊以较大优势击败了14名反对党候选人,再次当选为肯尼亚总统,任期5年。(孙,20)

区别词"现任"跟"前任"相对。好在这种跟职务变更动词同形的区别词不多。

3.2 职务变更动词中有一部分属于名动词,兼有名词的属性,表现为可以作主语和宾语、可以作定语直接修饰名词。[①] 例如:

(16) 人大常委会通过一批免职与任命名单。(南方网)

当这种职务变更动词作定语时,其论元角色不会出现,可以不加理会。

3.3 当职务变更动词的论元结构作宾语小句时,论元角色关系不变。例如:

(17) 墨西哥恰帕斯州议会7日批准[胡里奥·鲁依斯辞去州长职务],并任命众议员罗伯特·阿尔沃雷斯·纪廉为新州长。(孙,40)

(18) 坦桑尼亚总统姆卡帕5日在内罗毕祝贺[莫伊连任肯尼亚总统]时说……(孙,27)

(19) 咸阳市人大常委会$_i$ 决定[e_i 撤消张定会副市长职务]。(新华网)

(20) 他$_i$ 决定[e_i 辞去外长职务]。(孙,46)

(21) 十届全国人大常委会第五次会议$_i$28日下午通过表

① 关于名动词,详见朱德熙(1982)和朱德熙(1985)。

决，决定[e_i 任命孙文盛为国土资源部部长]。(中国网)

(22) 陕西省咸阳市第四届人民代表大会常务委员会第26次会议$_i$13日通过了{[关于(e_i 撤消张定会咸阳市人民政府副市长)]的}决定。(新华网)

(22)是“撤消”及其论元组成的述宾结构作介词“关于”的宾语，[①]然后加“的”作名词化转换；这个“的”字结构再作“决定”的定语，最后这个复杂的NP作动词“通过”的宾语。四层嵌套之后，“撤消”及其论元的论旨角色和句法分布保持不变。

必须注意的是，当担任动词的前面有“接替”等动词、从而构成“NP_1＋接替＋NP_2＋担任＋Re”一类格式时，就表示“NP_1”担任了Re，同时，“NP_2”辞去了Re。例如：

(23) 曾庆红接替胡锦涛兼任中央党校校长。(南方网)

也就是说，“接替”类动词可以使担任动词句同时表示任职和辞职两个事件。

3.4 职务变更动词跟其论元组成的主谓结构之间可以插入助词“的”，来使整个结构名词化；但是，论元角色关系保持不变。例如：

(24) 鲁依斯的辞职受到各方面欢迎。(孙，40)

因此，在信息抽取时可以不理会这种职务变更动词之前的“的”。

3.5 当职务变更动词的论元结构后附助词“的”表示自指(self-designation)时，论元角色关系不变；这种“的”字结构之后一定有NP作中心语。例如：

(25) [中国政府撤销卫生部长张文康职务]的决定是完全正确的。(联合早报)

(26) [贾安庆被撤职]的原因是……(中新网)

(27) [由托绍夫斯基出任总理]的新政府共有十八名成员，……(孙，10)

(28) 1939年11月，成立了[以薄一波同志任书记]的晋东

① “关于”能带动词性成分或小句作宾语，详见吕叔湘(2001)，第240页。

南军政委员会。(孙,28)

(29) 陕西省体育局今天对陕西体(育)彩(票)中心主任贾安庆作出[撤职]的决定。(中新网)

如例(27)(28)所示,当这种作中心语的NP是机构名词时,正好是要抽取的Org(组织)这种信息项目。例(29)则是更为复杂的表达方式,"撤职"的经事隐藏在主句的状语中。

3.6 当职务变更动词的论元结构后附助词"的"表示转指(transferred-designation)时,虽然论元角色关系不变,但是论元的位置会发生有规律的变化。这种"的"字结构之后作中心语的NP肯定是动词的必有论元。例如:

(30) a. 省委常委、常务副省长汪洋曾任省体委主任

→ b. [曾任省体委主任]的省委常委、常务副省长汪洋提高了嗓门……(孙,19)

→ c. *省委常委、常务副省长汪洋曾任的省体委主任

(31) 冈崎嘉平太,这位[曾任日本全日空航空公司总裁]的老人,生前一百多次来中国为恢复中日邦交奔波。(孙,36)

(32) a. 拉菲克·塔拉尔新当选巴基斯坦总统

→ b. 新当选巴基斯坦总统的拉菲克·塔拉尔

→ c. [新当选]的巴基斯坦总统拉菲克·塔拉尔今天在这里宣誓就职。(孙,13)

(33) [新当选]的中央政治局常委与中外记者见面。(南方网)

(34) [在肯尼亚大选中赢得连任]的肯尼亚总统莫伊5日在内罗毕宣誓就职。(孙,20)

(35) a. 上级刚任命童志成当一把手

→ b. 公司其他领导……,也都不约而同地来见见[上级刚任命当一把手]的童志成。(孙,37)

虽然"的"既可以提取主语,又可以提取宾语,但是,用在担任动词之后时,一般只能提取主语,不能提取宾语(如30b,c)。"新当选的"在句法、语义功能上相当于区别词"新任",后面一般是由组织、职务名

词和人名构成的同位性偏正短语（如32c），或者直接由组织、职务名词称代担任这种职务的人名（如33）。这种“的”字结构后面的组织、职务、人名正好是要抽取的信息项目。

4 用篇章结构知识找回缺失的信息项目

在有的职务变动文本中，信息抽取模型的四个模板元素（人物、时间、职务、组织），不一定在职务变更动词所在的小句中都出现。于是，就需要利用篇章结构知识来求解这些缺失的、或以代词、指示词(deixis)形式表达的信息项目。

4.1 利用语篇中的话题，求解以代词或空语类形式出现的经事（人物）。例如：

(36) 利维$_i$当天下午在特拉维夫宣布，{由于内塔尼亚胡$_j$未能对他所提出的一些修改1998年度国家预算的要求作出答复，他$_i$决定辞去外长职务}。(孙，47)

(37) 据报道，该暗杀团伙的主谋$_i$是哈马·阿马杜$_i$。他$_i$现任尼（日尔）反对派“社会发展全国运动”总书记，[e_i]过去曾担任过政府总理。(孙，22)

(38) 邢云$_i$，1952年生，大学文化。[e_i]历任内蒙古伊克昭盟副盟长、盟委副书记、盟长，[e_i]1996年10月起任盟委书记、盟人大工委主任。(孙，23)

(39) 列希$_i$曾连续担任过十一议员，[e_i]并在一九五三年至一九八二年期间担任阿尔巴尼亚人民议会主席团主席国家元首之职。(孙，15)

(40) 现年60岁的艾赛义德$_i$是摩（洛哥）右翼党派宪政联盟的议员，法学博士，[e_i]曾担任过国务秘书和阿拉伯议会联盟委员会主席等职。(孙，34)

从上例可以看出，这种代词或空语类的先行语(antecedent)都是前面小句中作主语（或主语中的中心语）的人名，特别是具有高话题性的

人名。[①] 比如,(36)中"他"最靠近的人名是"内塔尼亚胡",但是整个语篇的话题是主句主语"利维"。(37)的后续句中的"他"也跟先行句的主语同指,但是这个主语的具体所指要靠其后的同指宾语来确定。在我们调查的81个职务变动句中,这种以代词或空语类作经事的有13句,约占16%。基本上都能用先行句的主语来求解其所指。据此,可以得出规则:作经事的代词或空语类的先行语就是先行句中的人名,特别是句首具有高话题性的人名。

4.2 利用先行句中的时间,求解缺失或以指示词表达的时间。例如:

(41) 据新华社伊斯兰堡1月1日$_i$电(记者杨士龙)新当选的巴基斯坦总统拉菲克·塔拉尔今天$_i$在这里宣誓就职。(孙,13)

(42) 新华社斯德哥尔摩1月7日$_i$电(记者许福瑞)曾为调解巴勒斯坦和以色列冲突作出过努力的挪威政府,最近$_i$任命罗德—拉森为驻中东巡回大使……(孙,31)

(43) 塔拉尔于1997年$_i$3月当选巴参议院议员,同年$_i$12月15日被执政党穆斯林联盟谢派提名为总统候选人。(孙,11)

(44) 何长工1952年8月$_i$调入地质部。此前$_i$曾任重工业部副部长、代部长……(孙,29)

(45) 1984年,广昌成立了全国第一个白莲科研所,刘光亮担任所长……(孙,6)

(46) 大连市第十二届人民代表大会第一次会议1月10日选举于学祥为市人大常委会主任,选举薄熙来为大连市市长。(孙,39)

(47) 新华社北京1月6日$_i$电 中华人民共和国主席江泽民根据全国人民代表大会常务委员会的决定,任命王学贤为中华人民共和国驻南非共和国特命全权大使。(孙,26)

(48)"黎明俱乐部"共有18名成员。在成立大会上,白滨

---

① 至于高话题性(high topicality)到底有哪些形式指标,限于篇幅,暂不讨论。

一良被选为代表，木庭健太郎被选为干事长。(孙,9)

像“今天、同年”和“最近、此前”等是指示词，它们的所指要参照说话的情景或上下文才能确定；因此，又叫索引词语(indexical expression)。其中，前者是语义明确的指示词，指示某个特定的时间；后者是语义模糊的指示词，指示大概的某段时间。如例(41)—(44)所示，它们的参照词可以从先行句(包括报导引语)中找到。相应地，在信息抽取的输出文档中，这种时间应该采用“指示词—参照词”这种双重标记。比如：“今天—1月1日、此前—1952年8月”。对于未出现时间词语的句子，其时间一般可以在邻近的先行句中找到。如例(41)—(44)所示。像(48)这种句子，要在更先前的句子中找到成立大会的时间来确定任职的时间。必须注意的是，§2.4中(3)—(8)这类有标记的“曾任”事件，作者可能无意给出明确的时间。因此，碰到这类句子，如果本句中没有时间词语，那么可以不再求解。即把该时间搁置起来，处理为隐性的(covert)信息项目。据此，可以得出规则：如果本句中没有时间词语，那么其时间跟先行句中的时间一样；如果本句中有时间指示词语，那么其参照时间就是先行句中的时间。

4.3　利用先行句中的职务名词，求解本句中未出现的职务。例如：

(49) 陕西省体育局今天对陕西体(育)彩(票)中心主任贾安庆作出撤职的决定。(中新网)

(50) 陕西体(育)彩(票)中心领导班子被勒令辞职。(南方网)

(51) 新当选的巴基斯坦总统拉菲克·塔拉尔今天在这里宣誓就职。(孙,13)

(52) 以克劳斯为首的捷克原政府是一九九七年十一月三十日被迫辞职的。(孙,10)

(53) 墨西哥恰帕斯州议会7日批准胡里奥·鲁依斯辞去州长职务……鲁依斯的辞职受到各方面欢迎。(孙,40)

(54) 墨西哥总统塞迪略任命女参议员罗萨里奥·格林为外交部长。格林同日在就职后宣布……(孙,35)

职务名称及其所属的组织名称，是及物的职务变更动词的必有论元，即系事 Re，一般情况下是必须出现的。像“撤职、辞职、就职”等不及物动词，在句法上不允许带 Re 这一论元，于是就以经事的修饰语或动词的状语等形式出现（如 49—52）。只有在后续小句中以内嵌或名词化形式出现时，才可以承上省略（如 53、54）。值得注意的是，职务可以通过转喻（metonym）的修辞手法而借代担任这种职务的人，于是造成经事（实体）跟系事（属性）的合一（如 50、52、33）。据此，可以得出规则：如果动词之后没有职务名词，那么到动词之前去找；如果本句中没有职务名词，那么到先行句中去找。

4.4 利用先行句中的组织名词，求解本句中未出现的组织。例如：

(55) 香港特区政府昨天公布了香港特区基本法推广督导委员会成员名单，政务司长陈方安生出任委员会主席，高苕华任副主席。（孙，2）

(56) 1938 年 7 月，在沁县办事处的基础上成立了中共太岳特委，安子文同志任书记……。1939 年 4 月，中共太岳特委改称中共太岳地委，安子文同志任书记。（孙，28）

(57) 国务院调整三峡工程建设委员会，温家宝兼任主任。（南方网）

(58)“黎明俱乐部”共有 18 名成员。在成立大会上，白滨一良被选为代表，木庭健太郎被选为干事长。（孙，9）

像“成立、改称、调整”动词后的组织名称，往往是后续的担任动词的系事所属的组织。例(58)是组织名称作主语这种具有高话题性的成分，因而后续句中可以省去对这个组织的交代。据此，可以得出规则：如果本句中没有组织名词，那么到先行句中去找。

5 结语：论元结构知识的广泛适用性

完整的信息抽取包括三个层次的任务：(i) 模板元素任务，抽取文本中相关的命名实体，诸如专有名词、时间词语、数量词语等；(ii)

模板关系任务，抽取命名实体之间的各种关系（事实）等，诸如 Location_of，Employee_of，Product_of 等关系；(iii) 脚本（scenario）模板任务，抽取指定的事件，包括参与这些事件的各个实体、属性或关系；比如，航天器发射事件及其涉及的运载工具、负载客体、时间和场地等。[①] 像孙斌(2000)的 InfoX，就是一个脚本模板技术模型。

显然，动词的论元结构是一种非常适合于上述任务的语言知识。论元角色对应于模板元素，论元之间的论旨角色关系对应于模板关系，相关的一组动词的论元结构及其关联对应于脚本模板。再辅之以基于论元结构的逻辑结构和篇章结构知识，那么信息抽取就获得了比词语切分、词类、短语边界和句法成分等更具结构性的、更针对工作目标的语言知识。汉语语法学界自上世纪八十年代以来，在动词配价研究名目下对汉语动词的论元结构进行了大规模的研究，这些研究成果值得从事信息抽取研究的学者去改造和利用。

参考文献

吕叔湘主编（2001）《现代汉语八百词》，北京：商务印书馆。

孙　斌（2000）《继承—归纳机制及其在对象系统中和信息提取技术中的应用》，北京大学计算机系博士学位论文。

袁毓林（1998）《语言的认知研究和计算分析》，北京：北京大学出版社。

袁毓林（2002）《信息抽取的语义知识资源研究》，《中文信息学报》第5期。

袁毓林（2005）《用动词的论元结构跟事件模板相匹配》，《中文信息学报》第5期。

朱德熙（1982）《语法讲义》，北京：商务印书馆。

朱德熙（1985）《现代书面汉语里的虚化动词和名动词》，《北京大学学报》第5期；收入《语法丛稿》，第114—124页，上海：上海教育出版社。

2004年6月初稿，2004年9月改定

（发表于《中文信息学报》2005年第4期）

---

① 详见孙斌(2000)，第105页。

基于论元结构的
语义标注的体系和规范

本文讨论对汉语真实文本进行语义关系分析和标注的体系和规范，说明篇章语义、论旨角色和逻辑语义这三种层面的语义关系，都能以论元结构为基础来进行分析和标注，从而提出了一种基于论元结构的汉语语义关系的标注体系：(i) 以谓词的论旨结构为基础，给谓词所支配的各个论元标注论旨角色；(ii) 给附加在论元结构上的否定、时体和模态算子等逻辑成分标注语义功能及其辖域，给指代词标注照应关系；(iii) 给联结不同的论元结构的语篇衔接词语标注篇章功能及其配对关系。同时，为这三个层面上的各种语义关系设计了便于记忆的标记，形成了一套可扩充的标记集。并且，还为各种语义关系的标注制定了比较具体的操作规范。最后，展示怎样从经过上述语义标注的语料上自动地为句子建立语义依存树和句法关系树，还讨论了这种语料库在信息抽取、机器学习等领域的应用。通过对数万字新闻文本的手工标注，显示出这套标注体系对真实语料具有较好的适应性和较高的语义关系信息的覆盖率。

1 语义标注的目标、内容和体系

袁毓林(2002)指出，为了给信息抽取(information extraction)等自然语言信息处理提供充分的语义知识方面的资源，有必要对一定数量的真实文本进行语义关系标注，建立一种标有语义关系网络的精炼语料库。[①] 要进行语义关系标注，首先要解决的问题是标注什么，即标注的内容问题；其次要解决的问题是怎么标注，即标注的规范问题。其实，标注什么和怎么标注，是受语义标注的目的引导

---

① 为了区别于树库(tree bank，即标注了句法树关系的语料库)，标注了语义网络关系的语料库可以叫作网库(net bank)。

的。我们以从真实文本中抽取用户指定的信息为应用目标，通过一段时间的尝试后发现，对于信息抽取等广泛地涉及到词项之间和句子之间的语义关系的语言信息处理工作来说，如下三个层面的语义知识是非常重要的：(i) 篇章结构关系，包括小句之间、句子之间、甚至段落之间的语义关系；(ii) 论元结构关系，主要是动词、形容词等谓词性成分跟受其支配的体词性成分之间的语义关系；(iii) 逻辑结构关系，主要是否定算子、时体算子、模态算子跟受其约束的成分之间的逻辑语义关系，也包括代词、指示词(deixis)跟其先行语之间的照应关系(anaphoric relation)。

通过对职务变动文本的语义关系分析，袁毓林(2005a，b)又发现：上述三个层面的语义知识可以以论元结构知识为核心来组织和表示。这样，逻辑结构便是附加在以动词为中心的论元结构之上的否定、时体和模态算子跟论元结构中的有关成分的语义约束关系；篇章结构便是从前面的论元结构到后面的论元结构的推进和关联，其中也涉及到论元的传递、称代和省略等问题。对这些内容进行标注，就形成了基于论元结构的语义标注体系。

有了这些理论和认识上的准备，我们开始对关于职务调动的真实新闻文本进行语义关系标注实践，逐步摸索出一套简明自然、大体完备自洽的基于论元结构的语义标注体系及相关规范。下面，分别按上述的三个层面分别介绍和讨论。

2 论旨角色关系的标注及其规范

2.1 论旨角色的名称和定义

动词的论元结构包括动词的论元属性(可以支配的论元数目)、论旨角色关系(这些论元跟动词的语义关系)、配位关系(动词跟其论元可以构成哪些句法格式)等内容。对于在真实文本上进行语义标注来说，只有论旨角色关系才是必须标明的语义信息，其他暂时撇开不管。论旨角色关系最终体现为给受动词支配的论元指派施事、受事等语义角色，简称论元角色。关于论元角色的种类、名称、定义、及

其英语速写，我们基本按照袁毓林(2002b)。根据不同论元跟动词的语义关系和句法实现的情况，可以把论元分为必有的(obligatory)和非必有的(non-obligatory)两大类，必有论元又分为主体论元(subject argument)和客体论元(object argument)两小类；非必有论元又分为凭借论元(means argument)和环境论元(environment argument)两小类；然后是各种具体的论元，当然底下还可以再分出各种小类来。

从我们对真实新闻文本的语义标注实践来看，经常碰到的是以下这些论元：

(一) 必有论元：

A. 主体论元：

(1) 施事(agent，简写为 A)：自主性动作行为的施行者。

(2) 感事(sentient，简写为 Se)：非自主性的心理感觉的主体。

(3) 经事(experiencer，简写为 Ex)：某种变化的具有感知性的主体。

(4) 致事(causer，简写为 Cau)：某种致使性事件的引起者。

(5) 主事(theme，简写为 Th)：性质、状态等无施动、感知性的主体。

B. 客体论元：

(1) 受事(patient，简写为 P)：因施事的行为而受到影响的事物。

(2) 与事(dative，简写为 D)：动作、行为的非主动的参与者。

(3) 结果(result，简写为 R)：动作、行为造成的结果。

(4) 对象(target，简写为 Ta)：感知性动作、行为的对象和目标。

(5) 系事(relative，简写为 Re)：事件中跟主体论元相对的其他各种客体。

(二) 非必有论元

A. 凭借论元：

(1) 工具(instrument,简写为 I)：动作、行为所凭借的器具。

(2) 材料(material,简写为 Ma)：动作、行为所用的材料。

(3) 方式(manner,简写为 M)：动作、行为所采取的方式、方法。

(4) 原因(reason,简写为 Rn)：动作、行为、事件等发生的原因。

(5) 目的(aim,简写为 Ai)：发生动作、行为、事件等的目的。

B. 环境论元：

(1) 时间(time,简写为 T)：动作、行为、事件等发生的时间。

(2) 处所(location,简写为 L)：动作、行为、事件等发生的处所。

(3) 源点(source,简写为 So)：动作、行为、事件等开始的时间或处所。

(4) 终点(goal,简写为 Go)：动作、行为、事件等结束的时间、处所或状态。

(5) 路径(path,简写为 Pa)：动作、行为、事件等中途经过的时间或处所。

(6) 范围(range,简写为 Ra)：动作、行为、事件等所涉及的数量、频率、幅度、时间等事项。

事实上,源点、终点和路径通常是跟处所相关的,于是,我们约定：源点处所记作 L(S),终点处所可以记作 L(G),介于源点和终点之间的路径处所可以记作 L(P)。

2.2 论旨角色的标注规范

(1) 在原则上,论旨角色关系的标注是以动词为中心的,假定每一个动词(特别是作谓语核心的动词)都构成一个论元结构。于

是,(承上下文省略的主体论元、客体论元等必有论元都看作是空语类(empty category),用[e]作标记。并且,在这个空语类和其先行语上加同指(coreference)下标(依次为 i, j, k,…);当空语类的先行词不止一个(即空语类是复数形式)时,空语类之后的几个下标用加号连接;当空语类的先行语不明确时,可以用问号作下标。(对于隐含在语境中的必有性论元成分,用PRO(大代号)作标记。例如:

a. {1989 年}$_T$[[梁惠珍$_i$]$_{Ex}$退休后]$_T$,[[e_i]与丈夫$_j$]$_A$ 回到[老家湛江]$_{L(G)}$,[e_j]$_A$ 开办〈了〉$_{perf}$[“惠珍联合医院$_k$”]$_R$,[e_k]专治[男女不育症]$_P$。

b. {经过几年奋斗}$_M$,[他$_i$]$_A$ 带领[乡亲们$_j$]$_P${通过股份合作制}$_M$,[e_{i+j}]$_A$ 办起[集农、科、贸一体的农业集团公司]$_R$,[e_{i+j}]$_A$ 实现〈了〉$_{perf}$[农业产业化]$_R$。

c. {[他$_i$]$_{Ex}$ 上任后}$_T$,[PRO_j]$_A$ 立即召开[党委会]$_R$,[PRO_j]$_A$ 研究[(跟)群众息息相关的治安问题]$_P$。

d. [邱娥国$_i$ 的职务]$_{Ex/Th}${虽}$_{CES\text{-}i}$〈已〉$_{past}$ 升为[分管户籍、外勤的副所长]$_{Re}$,{但}$_{VER\text{-}i}$[他$_i$]$_A$ 还是{“按照原来的那样}$_M$ 做[PRO]$_{P/R}$”,……。

一般地说,空语类通常是可以根据上下文明确地补出来的,当然由于句法结构方面的限制,补出来的语句形式不一定是合语法的。大代号一般没有先行语,或者先行语不明确。比如,上例 c 中“召开党委会,研究……”的应该是“李常水”和“党委一班人”。

(2) 对于一个句子中有多个动词性成分,分为下面几种情况:① 对于叙述同一个主语的一连串小句,为每一个动词性成分标注其论元角色,省略的用空语类补充出来。如果连动式中间用逗号断开,那么看作不同的小句。② 对于几个动词连用构成的并列结构(如“指导和协调”)、关系紧密的连动式(如“报道说、宣誓就职”)、述补结构(“开进、组装成”)、“形式动词+名动词”组合(如“作斗争”)、熟语性动词组合(如“说好话、感兴趣”)和动词重叠形式(如“读一读”)等动词性结构,把它当作一个谓词来标记其从属成分的论旨角色。③

对于一般所说的兼语式（比如“使”字句），在兼语成分上同时标记其相对于前后两个动词性成分的论旨角色，用合取符号 & 来连接这两个论旨角色标记。① 同样，对于表示存在的“有＋NP＋VP”等格式，根据 NP 跟“有”和 VP 的语义关系，分别标注两种论旨角色。④ 对于内嵌小句（比如“报道”等动词所带的宾语小句），也需要进行语义标注；通过加括号来标志其嵌套层次，不同层次的括号后面加上不同层次的论旨角色名称。⑤ 为了语义标注的精细，“的”字结构用圆括号标示，其中的谓词性成分，可以标记其论元角色等语义关系，其中跟中心语同指的空语类用下标来标示。当然，为了减少标注的层次，也可以暂时把“的”字结构当作一个 NP 而不作语义关系分析和标记。例如：

a. {1989 年}$_T$[[梁惠珍$_i$]$_{Ex}$退休后]$_T$，[[e$_i$]与丈夫$_j$]$_A$ 回到[老家湛江]$_L$，[e$_j$]$_A$开办〈了〉$_{perf}$[“惠珍联合医院$_k$”]$_R$，[e$_k$]专治[男女不育症]$_P$。

b. [他$_i$]$_A${先}$_{TEM}${从日本}$_{L(S)}$ 进口[原装马自达 323$_j$]$_P$ 至[香港]$_{L(G)}$，[e$_i$]$_A$ 就地拆散[e$_j$]$_P$，{按配件}$_M$ 报关进口。

c. [许多人$_k$]$_A$ 便主动找上[门]$_L$ 来，[有的$_{k1}$]$_A$ 拉[他$_i$]$_{P\&A}$合伙做生意，[有的$_{k2}$]$_A$〈想〉$_{mod}${[找[他$_i$]$_{P\&A}$做靠山]$_{Re}$}。

d. [俄罗斯国家杜马（议会下院）$_i$]$_A${6 月 11 日}$_T$ 出台[一项法律]$_R$，[e$_i$]$_A$ 授权[国防部]$_{D\&A}$ 掌管[武装部队的重要军事行动]$_{Re}$，[克瓦什宁的总参谋部]$_A$ 主要负责规划[俄罗斯未来的军事进程]$_R$。

e. [分析家]$_{Se}$认为，[[[PRO]$_A$ 指责[克瓦什宁]$_{P\&Ex}$[为印古什遇袭事件]$_{Re}$负责]$_{Th}$是有[一定依据]$_{Re}$的]$_{Re}$。

f. [中国卫生部常务副部长高强]$_A$ 说，[“[实践]$_{Th}$证明”，[（[中国政府]$_A$ 撤销[卫生部长张文康职务]$_{Re}$的）决定]$_{Th}$ 是

---

① 虽然从语法理论上讲，这会违反生成语法的 GB 理论中的论旨原则：一个论元只能担任一种论旨角色，一种论旨角色只能赋予一个论元。但是，从语言信息处理工程的角度上讲，这种合成标记法比较经济，也便于识别和处理，具有更高的效率。关于论旨原则，详见徐烈炯(1988)第 271 页。

["完全正确的"]$_{Re}$]$_{Re}$]$_{Re}$。

g. [冈崎嘉平太$_i$,这位([e_i]曾任[日本全日空航空公司总裁]$_{Re}$的)老人]$_A$,{生前}$_T$ 一百多次来[中国]$_{L(G)}$ {为恢复中日邦交}$_{Ai}$奔波。

(3) 对于省略的动词性成分,用空动词符号[V]作标记。例如:

a. [孙文盛$_i$]$_{Ex}$,男,[e_i]$_{Ex}$[V][61岁]$_{Re}$,[e_i]$_{Ex}$[V][山东威海人]$_{Re}$,[e_i]$_{Ex}$[V][大学学历]$_{Re}$,[e_i]$_{Ex}$[V][工程师]$_{Re}$。

b. {现在}$_T$,[十届全国人民代表大会]$_{Th}$ 实有[代表]$_{Re\&Th}$ [V][2977人]$_{Re}$。

c. [……现任总统丹尼尔·阿拉普·莫伊$_i$]$_A${以较大优势}$_M$ 击败〈了〉$_{perf}$[14名反对党候选人]$_P$,[e_i]$_{Ex}$再次当选为[肯尼亚总统]$_{Re}$,[任期]$_{Th}$[V][5年]$_{Re}$。

(4) "是……的"、"就是"等强调性成分、"被迫"、"畅行无阻"等方式性状语等成分暂不标记。例如:

a. [以克劳斯为首的捷克原政府]$_{Ex}$是{一九九七年十一月三十日}$_T$ 被迫辞职的。

b. [上苍]$_A$ 似乎故意地考验[她$_i$]$_P$。

(5) ① 为了区别,必有论元用方括号标志,非必有论元用花括号标志,并分别在这两种括号的后面标上论元角色的名称(首字母大写的标记)。② 同一种论旨角色,对于有的动词或动词短语来说可能是必有论元,但是对于另一些动词或动词短语来说则可能是非必有论元。(当对某个论元成分的论旨角色不能明确断定时,就把最可能的论旨角色名称依次都标上,不同的论旨角色名称中间用斜撇(/)来表示析取(disjunction)关系。例如:

a. [阿尔巴尼亚前人民议会主席团主席列希]$_{Ex}${一日晚}$_T$ {在地拉那}$_L$ 病逝,……

b. ……[龚德俊的这批"海马"]$_A$ 偷逃〈了〉$_{perf}$[关税、增值税、消费税]$_P$,畅行无阻地开进〈了〉$_{perf}$[京城]$_{L(G)}$。

c. [捷克总统哈维尔]$_A${二日}$_T$ 任命〈了〉$_{perf}$[捷克新政府成

员]$_{Ex/Re}$。

d. {([陈云峰]$_{A}$ 读[电大]$_{Re/M}$)期间}$_{T}$，…… {1994 年}$_{T}$，[黄赛红]$_{A/Ex/Se}$考取〈了〉$_{perf}$[浙江省政法管理干部学院]$_{P/Re/Ta}$，

e. [诸葛彩华]$_{Ex}$[从县委副书记岗位]$_{L(S)/Re1}$ 调任[代县长]$_{L(G)/Re2}$。

f. {在此期间}$_{T}$，[工商系统]$_{A}$ 实行 [体制改革]$_{P/M/R}$，……

在例 b 中，终点性处所“京城”是“开进”的必有论元。当“任命”等动词的客体论元是指人名词时，其论旨角色可以归入经事；当“任命”等动词的客体论元是职务名词时，其论旨角色可以归入系事；在例 c 中，“捷克新政府成员”似乎既涉及人员，又涉及职务；为了周全，可以把这两种论旨角色都标记上去。在例 d 中，对于动词“读”来说，“电大”既像是系事，又像是方式；对于动词“考取”来说，其主体性论元既像是施事、又像是经事或感事，其客体论元既像受事、又像系事或对象。对于动词“调任”来说，原来的职务和后来的职务都是系事；同时，从路径隐喻(path metaphor)的角度来看，原来的职务是源点，后来的职务是终点。

(6) 对于用介词引导的论元，我们约定：① 出现在动词之前时，整个介词结构都置于一个括号中，即作为一个论元成分；出现在动词之后时，把动词和其后的介词看作一个动词性成分，把介词之外的论元成分置于一个括号之中。这样，可以方便地处理这种结构中的时态助词(如“回到了〔故乡〕、埋在了〔城外〕”)。[①] ② 介词是论元角色的标志，俗称“格标记”(case marker)。为了醒目，跟动词一样，介词也用着重点标注。③ 为了一致和醒目，动词之后跟动词不连续的介词也放在表示论元成分的方括号之外；比如，把“任命……为”等看作是一个动词性结构，即一种不连续的动词性成分。④ 当“被”等引导必有论元的介词之后不出现宾语时，应该用空语类[e]作标记，并加上同指下标和论旨角色标记。这种带空语类作宾语的介词，因为作为动词短语的一部分，所以不用放在方括号中，即单独把空语类放在

---

① 这样处理有句法、语义和音系学上的考虑，详见袁毓林(2003)。

括号中。⑤ 承上文而省略的成分"被……"等引导必有论元的介词结构,用双重方括号中[[被 e]]作标记,并给空语类标记 e 加上同指下标,和在[被 e]之后加上论旨角色标记。例如:

a. [新当选的巴基斯坦总统拉菲克·塔拉尔]$_{A}${今天}$_{T}${在这里}$_{L}$ 宣誓就职。

b. [格林]$_{Th}${1941 年}$_{T}$ 生于[墨西哥城]$_{L}$。

c. [大连市第十二届人民代表大会第一次会议]$_{A}${1 月 10 日}$_{T}$ 选举[于学祥]$_{Ex}$为[市人大常委会主任]$_{Re}$,选举[薄熙来]$_{Ex}$为[大连市市长]$_{Re}$。

d. [塔拉尔$_{i}$]$_{Ex}${于1997 年$_{j}$3 月}$_{T}$ 当选[巴参议院议员]$_{Re}$,[e_i]$_{Ex}${同年$_{j}$12 月 15 日}$_{T}$[被执政党穆斯林联盟谢派]$_{A}$ 提名为[总统候选人]$_{Re}$。

e. [白滨一良]$_{Ex}$被[$e_?$]$_{A}$ 选为[代表]$_{Re}$,[木庭健太郎]$_{Ex}$被[$e_?$]$_{A}$ 选为[干事长]$_{Re}$。

(7) 为了简单,把邻接的复指性成分处理成一个论元。例如:

[冈崎嘉平太$_{i}$,这位([e_i]曾任[日本全日空航空公司总裁]$_{Re}$的)老人]$_{A}$,{生前}$_{T}$一百多次来[中国]$_{L}${为恢复中日邦交}$_{Ai}$奔波。

(8) 不同的句法分析,可能导致不同的语义标注。应该从中选择反映语义关系最明确的一种标注;在其他情况相同的条件下,尽可能选择相对简单的标注。例如:

a. [十届全国人大常委会第五次会议$_{i}$]$_{A}${28 日下午}$_{T}${通过表决}$_{M}$,决定[[e_i]$_{A}$ 任命[孙文盛]$_{Ex}$为[国土资源部部长]$_{Re}$]$_{Re}$。

a'. [十届全国人大常委会第五次会议$_{i}$]$_{A}${28日下午}$_{T}$ 通过[表决]$_{P}$,[e_i]$_{A}$ 决定[[e_i]$_{A}$ 任命[孙文盛]$_{Ex}$为[国土资源部部长]$_{Re}$]$_{Re}$。

b. [十届全国人大常委会第五次会议$_{i}$]$_{A}${28 日}$_{T}${经表决}$_{M}$ 通过[决定]$_{P}$,[e_i]$_{A}$ 免去[田凤山的国土资源部部长职

务]$_{Re}$;[e_i]$_A$ 任命[孙文盛]$_{Ex}$为[国土资源部部长]$_{Re}$。

c. {[$e_?$]$_A$ 违规调人}$_{Rn}$,[咸阳市人民政府副市长张定会]$_{P/Ex}$[被人大]$_A$ 撤消[职务]$_{Re}$。

c'. {[$e_?$]$_A$ 违规调人}$_{CAS}$,{[咸阳市人民政府副市长张定会]$_{P/Ex}$[被人大]$_A$ 撤消[职务]$_{Re}$}$_{CSQ}$。

d. [两名全国人大代表$_i$]$_P${因[e_i]$_A$ 涉嫌[贿选和收受贿落]$_{Re}$}$_{Rn}$被[$e_?$]$_A$ 罢免。

通过跟例 b 比较,可以发现:例 a 中的"通过",如果分析为介词,那么"通过表决"就是方式论元,整个句子是简单句;如果分析为动词,那么"十届全国人大常委会第五次会议 28 日下午通过表决"就是一个小句,整个句子是由两个小句构成的复合句,后一小句承前省略了主语。通过跟例 d 比较,可以发现:例 c 中的"违规调人"既可以分析为原因论元(整个句子是单句),也可以分析为原因小句后面的小句就是结果小句,整个句子便是复句)。相对来说,简单句比复合句简单,单句比复句简单。因此,优先考虑 a 和 c 这两种标注方式。

3 逻辑语义关系的标注及其规范

3.1 逻辑语义关系的种类和相关词语

根据上文 §1 的说明,逻辑语义关系是依附在论元结构之上的否定关系、模态关系、时体关系、称代关系和指示关系,主要涉及否定算子、模态算子和时体算子跟受其约束的成分之间的逻辑语义关系,还有代词和指示词跟其先行语之间的照应关系。可以分述如下:

(1) 在现代汉语中,否定算子(negative operator,简写为 neg)主要是副词"不"和"没、没有"。为了方便,助动词"别、甭"也可以算进去。在书面性较强的文体中,有时会用到"未"等副词。当然,文言色彩较重的"弗、勿、毋、莫"等,偶尔也会用到。

(2) 模态算子(modal operator,简写为 mod)主要是表示情态的

助动词,常见的有"能、能够、可以、会、可能、得(dé)、敢、肯、愿意、情愿、乐意、想、要、当、应、该、得(děi)、应该、应当、许、准、值得、配"等。副词"必须、一定"等也可以算进去。

(3) 时体算子包括"将、刚、刚刚、已经、曾经、又、再、正、在、正在"等时间副词、"着、了、过"等时态助词、"了、呢、着呢、来着、来的"等语气词。其中,"将、即将、再"等表示将来时(future tense,简写为fut),"刚、刚刚、已经、曾经、又"等表示过去时(past tense,简写为past);"着、呢、着呢、正、在、正在"表示进行体(progressive aspect,简写为prog),助词"了"和语气词"了"表示完成体(perfect aspect,简写为perf);"来着、来的"最近的过去发生的事,相当于现在完成体(present perfect,简写为pres-perf),"过"表示经历体,相当于过去完成体(past perfect,简写为past-perf)。

(4) 代词包括"我、咱、你、您、他(她、它)、我们、咱们、你们、他们(她们、它们)、大家、自己"等人称代词。

(5) 指示词包括"这、那、这里(这儿)、那里(那儿)、这么、这样、那么、那样、这么样、那么样"等指示代词,文言词"此"和包括"此"的词语也可以算进去。另外,像"昨天、今天、明天、此前、此后、此外、同年、同日、同时"等相对性时间词,它们不表示绝对的时间,只有根据话语中某个具体的时间来确定其所指,也属于指示词。

3.2　逻辑语义关系的标注规范

我们约定,表示逻辑语义关系的词语都用尖括号套起来,在括号外加上语义功能的标志。这一层面的标志,一律用小写字母组成的简称。具体地说,约有六项内容:

(1) 在否定算子之后一律加上 neg 作标志,并用花括号标志其辖域。例如:

a. [军队人数]$_{Th}$〈不〉$_{neg}${〈得〉$_{mod}${超过[全国人口总数的1%]$_{Re}$}}。

b. 偏偏{1995下半年}$_{T}${玉环岛上}$_{L}$[滴水]$_{Th}$〈未〉$_{neg}${下}

……

(2) 在模态算子之后一律加上 mod 作标志，并用花括号标志其辖域；这样，不仅标明了其所支配的动词性成分的范围，而且还标明了花括号中的语言表达在情态上表示的是一种非现实的断言(irrealis assertion)。例如：

a. [领导]$_{A}$〈要〉$_{mod}${[与运动员、教练员]$_{D}$ 滚在[一块]$_{L(G)}$……}

b. ……〈可能〉$_{mod}${〈会〉$_{mod}${下达[撤换其职务的命令]$_{P}$}}。

(3) 在各种时体算子之后，分别加上不同的时体标记。例如：

a. {前几天}$_{T}$，[我]$_{A}$ 还看望〈了〉$_{perf}$[张文康先生$_{i}$]$_{P}$……

b. [俄国防部和总参谋部]$_{A}${在过去 10 年中}$_{T}$，〈一直〉$_{prog}$〈在〉$_{prog}$激烈争夺[军事行动和军队经费掌管权]$_{P}$。

c. [列希]$_{Ex}$〈曾〉$_{past}$连续担任〈过〉$_{past\text{-}perf}$[十一届议员]$_{Re}$……

(4) 在代词或指示词语之后加上下标(依次为 i，j，k，…)，并在其先行语(或参照性词语)之后加上相同的下标，来显示同指(coreference)或索引(indexing)关系。我们约定：① 当先行语是整个论元时，由于不会引起歧解，因而不用加下划线；② 当先行语嵌在一个词组之中时，为了明确，需要在先行语之下加下划线；③ 当先行语是一个超出一个论元范围之外的复杂词组或小句、甚至复句、段落时，为了辨认的方便，用花括号把先行语套起来，并在括号后面加上下标。④ 当先行语不止一个、代词或指示词是复数形式时，代词或指示词之后的几个下标用加号连接。⑤ 相反，当代词性成分的所指跟先行语只是一种部分关系时，依次在这种部分代词的照应下标后面加数字(1，2，…)作标记。⑥ 对于先行语不明确的代词或指示词，用问号作下标。例如：

a. [博塔$_{i}$]$_{Ex}$〈曾〉$_{past}$任[旧南非的国防部长、总理和总统]$_{Re}$，[他$_{i}$ 掌握的重要情况]$_{Th}${对真相委员会完成其使命}$_{Re}$至关重要。

b. {1940 年$_{j}$1 月}$_{T}$，[中共北方局]$_{A}$ 决定[[太岳地区$_{i}$]$_{Th}$成

为[独立的战略区]$_{Re}$，[e_i]$_A$ 成立[太岳区党委]$_R$，[安子文同志]$_{Ex}$任[书记]$_{Re}$$_{Re}$。{同年$_j$1 月19 日}$_T$，[陈赓同志]$_A$ 奉[八路军总部命令]$_P$ 率领[八路军三八六旅]$_{P\&A}$进驻[太岳区]$_L$。

c. {[郴县工商局$_l$]$_{Th}${垂直上划过程中}$_T$，存在[([e_l]$_A$ 严重违规、[e_l]$_A$ 突击进人[70 多名]$_{Re}$和[领导干部]$_A$ 弄虚作假、以权谋私的)问题]$_{Re}$}$_m$，[对此$_m$]$_{Re}$[张定会]$_{Th}$负有[直接领导责任]$_{Re}$。

d. {在成立大会上}$_L$，[白滨一良$_i$]$_{Ex}$被[$e_?$]$_A$ 选为[代表]$_{Re}$，[木庭健太郎$_j$]$_{Ex}$被[$e_?$]$_A$ 选为[干事长]$_{Re}$。[他们$_{i+j}$]$_{Ex}$都〈曾〉$_{past}$担任[原公明党副书记]$_{Re}$。

e. {1995 年}$_T${([杨光$_i$]$_{Ex}$上任)后不久}$_T$，[许多人$_j$]$_A$ 便主动找上[门]$_L$ 来，[有的$_{j1}$]$_A$ 拉[他$_i$]$_{P\&A}$合伙做[生意]$_P$，[有的$_{j2}$]$_A$〈想〉$_{mod}${[找[他$_i$]$_{P\&A}$做[靠山]$_P$]$_{Re}$}。

在例 e 中，部分代词“有的”指人或事物中的一部分，其先行词是上文的“许多人”。两个“有的”的所指可以一样，也可以不一样；因此，在下标上加不同的数字以示区别。

(6) 为了跟先行语建立尽可能对等的同指或其他照应关系，可以把照应下标加在由代词或指示词所组成的词组后面，即整个这个词组跟先行语具有照应关系。例如：

a. [黄赛红]$_{A/Ex}$……，{于1996 年7 月$_k$}$_T$ 毕业。{也}$_{ADD}$就{在这个月$_k$}$_T$，[陈云峰$_i$]$_{Ex}$……调到[沙善办事处]$_{L(G)}$[e_i]$_{Ex}$当[副主任]$_{Re}$。

b. [鲁依斯的辞职$_j$]$_{Th}$受到[各方面欢迎]$_{Re}$。[一些人士]$_{Se}$认为，[[这样做$_j$]$_{Th}$〈可以〉$_{mod}${[对公正解决屠杀惨案，推动政府与该州游击队组织萨帕塔民族解放军恢复和谈]$_{Re}$，产生[积极影响]$_R$}]$_{Re}$。

为了使指示性词语跟先行语“1996 年 7 月”和“鲁伊斯的辞职”在语义照应关系上的一致，把照应标记分别加在整个指示性短语“这个月”和“这样做”之后。

4　篇章语义关系的标注及其规范

4.1　篇章语义关系和篇章关联词语

篇章语义涉及篇章中词、短语和句子语义之外的意义问题，主要有三个方面：(1) 话语的连贯，指前后句子之间在语义上的联系，包括前后命题之间的因果、条件、背景、限定、补充、解释等意义关系，在形式上则是通过语序、连词等语法或词汇手段来指示句子之间在语义上的联系。这是一种语篇的局部连贯(local coherence)。(2) 信息分布，指新旧信息在语篇中的表达方式和话语功能。话语的信息结构有话题和说明两个部分，前者表示已知信息，以此同上文相连接，后者表示新信息，以此推动话语的展开。它们可以在句子中表现为不同的句子成分，也可以在句群中表现为不同的句子。比如，话题在英语句子中可以由作主语的非重读的定指名词或代词担任，在句群中则可以由位于句群开头的从属小句担任。(3) 总体连贯(global coherence)，指语篇从头到尾的总体连贯，体现为语篇的宏观结构(macrostructure)。① 也就是说，篇章语义以意义连贯(coherence)为中心，要求表层篇章背后所指的篇章世界(textual world)中的每一个概念和关系都必须是相连的和相关的。这种意义连贯又是通过形式连贯(cohesion)来实现的，体现为通过各种形式、意义手段来使一个篇章中的各构成成分相互有联系；比如，词语的重现或部分重现、相同结构的并列、内容的复述、词语的称代和省略、时体成分、连接成分、“话题—说明”类信息结构、甚至特定的语调等超音段成分(suprasegment)。要而言之，一个语段之所以是篇章(而不是非篇章)就在于其有篇章组织性(texture)；而篇章组织性是建立在各种连贯关系(cohesive relation)上的，这连贯关系一方面使得对篇章中某些成分的解释需要依靠其他成分才得以进行，另一方面又是通过连接成

① 主要根据 van Dijk (1985)的有关见解，中文介绍详见陈平(1991)第 83—84 页。

分、指称、替代、省略和词义联系等来实现的。[①]

目前,语言学界对篇章语义的研究还不成熟,而且理论见解多所分歧。但是,根据上文§1的说明,篇章语义关系主要是骑跨在不同的论元结构之上的各种衔接关系(cohesion),诸如并列、选择、递进、连贯、转折、因果、假设、条件、目的、解释、承接、反意、总结等关系。主要涉及到各种关联词语及其所表示的篇章关系。另外,像话题—说明这种语用平面上的意义关系,对篇章组织和篇章语义也起作用。至于代词、指示词和省略等手段,当然对篇章组织和篇章语义也起作用;只是上文已经在放在论元结构层面上作了处理了,这里就不再重复。

在现代汉语中,表示小句或句子之间的语义关系的词语主要是连词、关联副词等。下面列出主要的篇章关系和相应的关联词语、及其缩写标记:

(一)承接性关系

(1)并列关系(coordinate,简写为COR),例如:"也、还、又、同时、同样、既……又……、一边……一边……"等。

(2)递进关系(additive,简写为ADD),例如:"还、进而、再说、再者、何况、况且、乃至、甚至、不但/不仅/不止/不光/不独/不单/非但/非特……而且/并且……"等。

(3)选择关系(alternative,简写为ALT),例如:"或、还是、或者……或者……、要么……要么……、不是……就是……、宁可/宁肯/宁愿……也(不)……、与其……宁可/宁肯/宁愿……"等。

(4)连贯关系(temporal,简写为TEM),例如:"首先……然后……接着……最后……"等。

(二)条件性关系

(5)条件关系(conditional,简写为CON),例如:"只要……就……、只有……才……、不管/不论/无论/任凭……都……、除非"等。

(6)因果关系(causal,简写为CAS),例如:"因此、因

① 主要根据de Beaugrande & Dressler (1981)和Halliday & Hasan (1976)、Brown & Yule (1983: 190—194)的有关见解,中文介绍详见廖秋忠(1992)第373—376、399、402页。

为……所以……、由于……因而……、既然……就/那么……”等。

(7) 假设关系(suppositional，简写为 SUP)，例如：“如果/假如/假使/要是……那么/就……、即使……也……、否则”等。

(8) 转折关系(adversative，简写为 VER)，例如：“而、却、反之、相反地、虽然……但是/可是/然而/不过……”等。

(9) 目的关系(purposive，简写为 PUR)，例如：“为了、以便、以免、省得”等。

其中，并列、递进、选择和连贯这四种关系，基本上表示几个命题之间的合取关系(conjunction)或析取关系(disjunction)，其中的每一个命题都是合取枝(conjunct)或析取枝(disjunct)可以简称为选枝(junct，记作 JUN)。而条件、因果、假使、转折和目的这五种关系，基本上表示两个命题之间的“条件—结果”关系，即蕴涵关系(implication)。比如，条件关系的前件表示条件(condition，简写为 CON)，后件表示结果(consequence，简写为 CSQ)。因果关系的前件表示原因(cause，简写为 CAS)，后件表示结果。假设关系的前件表示假设的条件(supposed condition，简写为 SUP)，后件表示结果。转折关系的前件表示让步性条件(concessive condition，简写为 CES)，后件表示结果；为了明确和区别，转折关系的后件标记为转折，即用 VER 作标志。目的关系由“手段—目的”两个命题组成，目的命题表示目的性条件(purposive condition，简写为 PUR)，手段命题(means，简写为 MEN)表示这种目的指导下的结果。

其他常用的语篇标志词语，如“例如、举一个例子、比方说、换一句话说、也就是说、如前所述、至于、总之、可见、一句话、显而易见、诸如此类”等，也可以归入连贯关系。

对于语义标注工作来说，对这种语篇关联词语(text conjunctives)及其表示的语义关系进行标注，是比较现实的。因此，我们把重点放在这一方面。

4.2 篇章语义关系的标注规范

(1) 为了显示语篇中的语义连贯关系，可以把副词“先、并”、连

词“然后、但是”等能把小句、句子、甚至段落等较大的语言单位连接起来，使之成为句子、话语、甚至篇章的词语或其他表层结构上的词汇特征，笼统地叫做话语衔接词语（cohesive expression，简写为COH）。所有的语篇标记，一律用大写字母。当话语衔接词语同时是论元性成分时，语篇功能标记加在论旨角色标记的后面，中间用合取号 & 连接。例如：

{1954 年$_{j}$}$_{T}$，[24 岁的梁惠珍$_{i}$]$_{Ex}$开始〈了〉$_{perf}$[行医生涯]$_{Re}$。[她$_{i}$]$_{A}${先}$_{TEM\text{-}i}$是{在一个县卫生所}$_{L}$工作，[e$_{i}$]$_{Ex}${后来$_{j}$}$_{T\&TEM\text{-}i}$成为[海南省屯昌县人民医院的妇产科主任]$_{Re}$。

像上例中的“后来”，既是时间论元，又是表示连贯关系的语篇衔接词语。

(2) 为了简单，对于单用的“先、并、也、但是、果真如此、总而言之、换句话说”等承接性语篇关联词语，用花括号套起来，并直接在其后标注其语义功能。当有关的语篇衔接词语的语篇功能不明确时，径直标上 COH；这一点，在其他地方也适用。例如：

a. {1994 年}$_{T}$，[黄赛红$_{j}$]$_{A/Ex/Se}$考取〈了〉$_{perf}$[浙江省政法管理干部学院]$_{P/Re/Ta}$，[e$_{j}$]$_{Ex}${通过三年的自费脱产学习}$_{M}$，{于1996 年 7 月$_{k}$}$_{T}$毕业。{也}$_{COR}$就{在这个月$_{k}$}$_{T}$，[陈云峰$_{i}$]$_{Ex}${因出色的工作成绩}$_{Rn}${从城关镇内设机构青马办事处}$_{L(S)}$调到[沙善办事处]$_{L(G)}$[e$_{i}$]$_{Ex}$当[副主任]$_{Re}$。

b. [南非真相委员会副主席伯瑞恩$_{i}$]$_{A}${同日}$_{T}$呼吁，[[他$_{i}$]$_{Se}$希望{[[博塔$_{j}$]$_{A}$〈能〉$_{mod}${{在最后一刻}$_{T}$改变[态度]$_{P}$，[e$_{j}$]$_{Se}$同意[[e$_{j}$]$_{A}$到[真相委员会]$_{L}$作证]$_{Re}$}]$_{Re}$}$_{k}$]$_{Re}$。{果真如此$_{k}$}$_{COH/SUP}$，[真相委员会]$_{A}$〈将〉$_{fut}$建议[[卡恩]$_{A}$撤诉]$_{Re}$。[博塔$_{j}$]$_{Ex}$〈曾〉$_{past}$任[旧南非的国防部长、总理和总统]$_{Re}$，[他$_{j}$掌握的重要情况]$_{Th}${对真相委员会完成其使命}$_{Re}$至关重要。[真相委员会]$_{A}$〈曾〉$_{past}$ 3 次传唤[[他$_{j}$]A 到场听证]$_{Re}$，{但}$_{VER}$[他$_{j}$]$_{A}$均{以有病等理由}$_{Rn}$拒绝[出席]$_{P}$，[e$_{j}$]$_{A}${并}$_{ADD}$指责[[真相委员会的工作]$_{Th}$是[“马戏团表演”和“政治迫害”]$_{Re}$]$_{Re}$。

在例 b 中，如果能肯定"果真如此"的语义功能是表示假设，那么标上 SUP；如果拿不准，就退一步标注其上位功能 COH。

(3) 对于"先……然后……、或者……或者……"等前后相呼应的表示承接性关系的关联词语，用花括号把成对的关联词语分别套起来，然后分别在花括号后面标上其语义功能，并加上相同的下标，以示它们之间的前后配套关系。例如：

[他$_i$]$_A${先}$_{TEM-i}${从日本}$_{L(S)}$进口[原装马自达 323$_j$]$_P$至{香港}$_{L(G)}$，[e$_i$]$_A$ 就地拆散[e$_j$]$_P$，{按配件}$_M$ 报关进口。{然后}$_{TEM-i}$，{再}$_{TEM-i}$[由广东、广西、湖北、四川等地的汽车修配厂]$_A$[将它们]$_P$ 组装成[车]$_R$。

(4) 对于"因为……所以……、如果……那么……、只要……就……"等前后相呼应的表示条件性关系的关联词语，或者(用花括号把成对的关联词语分别套起来，然后分别在花括号后面标注其语义功能，并加上相同的下标，以示它们之间的前后配套关系；或者(分别用花括号把其所关联的小句或句子套起来，然后分别在花括号后面标注条件(CON)和结果(CSQ)等语义功能，并在关联词语下面加着重点。例如：

a. 其实[大家]$_A${如果}$_{SUP-i}$认真地读一读[我在两次新闻发布会上所披露出来的中国卫生工作存在的各方面问题]$_P$，{就}$_{CSQ-i}$〈能〉$_{mod}${品味出[张文康工作中存在的失误]$_{Re}$}。

a'. 其实{[大家]$_A$ 如果认真地读一读[我在两次新闻发布会上所披露出来的中国卫生工作存在的各方面问题]$_P$}$_{SUP-i}$，{就〈能〉$_{mod}$品味出[张文康工作中存在的失误]$_{Re}$}$_{CSQ-i}$。

b. [邱娥国$_i$ 的职务]$_{Ex/Th}${虽}$_{CES-i}$〈已〉$_{past}$升为[分管户籍、外勤的副所长]$_{Re}$，{但}$_{VER-i}$[他$_i$]$_A$ 还是{"按照原来的那样}$_M$ 做[PRO]$_{P/R}$"，[e$_i$]$_A$ 经常深入[辖区]$_L$，[e$_i$]$_A${为实现辖区发案少、秩序好、群众满意}$_{Ai}$而努力。

b'. {[邱娥国$_i$ 的职务]$_{Ex/Th}$虽〈已〉$_{past}$升为[分管户籍、外勤的副所长]$_{Re}$}$_{CES-i}$，{但[他$_i$]$_A$ 还是{"按照原来的那样}$_M$ 做

[PRO]$_{P/R}$”，[e_i]$_A$ 经常深入[辖区]$_L$，[e_i]$_A${为实现辖区发案少、秩序好、群众满意}$_{Ai}$而努力}$_{VER\text{-}i}$。

第①种标记方案的优点是简洁，缺点是关联词语所领辖的小句或句子的界限不清楚；第②种标记方案的优点是关联词语所领辖的小句或句子的界限很清楚，缺点是碰到多重复句等复杂的句子或句群时，括号繁多，反而模糊了层次关系。

(5) 汉语中有一种无根话题句(dangling topic sentence)，[①]其句首话题跟说明部分中的谓语核心没有论元结构关系。显然，这种句子中的话题成分无法从说明部分中的谓语核心上得到论旨角色，结果使得我们无法标注其论旨角色。碰到这种情况，首先用花括号界定这两个成分的范围，然后分别在后面标注话题(topic，简写为 TOP)和说明(comment，简写为 COM)这种话语结构关系，最后在说明部分内部再标注其中的论元成分相对于谓语核心的论旨角色。例如：

a. {该剧}$_{TOP}${[情节]$_{Th}$曲折，[感情纠葛]$_{Cau}$让[人]$_{P\&Se}$回肠荡气}$_{COM}$。

b. {“宝马”假彩案}$_{TOP}$：{[陕西体彩中心主任贾安庆]$_P$被[e_j]$_A$撤职}$_{COM}$。

这样，可以让文本中的每一个名词性成分都能得到一个论旨角色或话语功能角色标记。

5　语义关系标注语料库的应用

5.1　从语义依存树到句法关系树

根据上述语义标注体系及其规范，我们就可以建造带有语义关系标记的语料库，即一种经过语义关系分析的语料库。由于语义关系是超越于树形结构的网络结构，因而这种标注了语义关系的语料

① 关于无根话题句，详见 Shi (2000)。

库应该是一种网库（net bank）。但是，如果着眼于谓词性成分跟其论元成分之间的论旨角色关系，舍弃其他方面的一些语义细节；那么，就可以从句子的论旨角色标记上自动地建立起一棵扁平形状的语义依存树（dependency tree）。其大概的方法是：(i) 首先建立句子节点 S，(ii) 然后把充当谓语核心的动词性成分提到直属于 S 的子节点 VP，(iii) 最后把一个个论元作为叶子节点连接到 VP 节点上。当然，论元成分中也可以包含动词性成分，其本身也可以是一棵依存树，从而形成依存树的递归结构。这种扁平形状的语义依存树还可以转换成有深度的句法结构树。其大概的方法是：(i) 首先，如果充当谓语核心的动词性成分之后有论元成分，那么把充当谓语核心的动词性成分跟紧邻其后的论元成分构成述宾关系；如果这个述宾结构之后还有论元成分，那么这个述宾结构再跟其后的这个论元成分构成复杂的述宾关系（双宾语结构）；(ii) 然后，从这个复杂的述宾结构依次向前，跟其前面各个无介词引导的论元成分渐次构成层层嵌套的主谓关系，跟其前面各个有介词引导的论元成分渐次构成层层嵌套的偏正关系。再考虑得复杂一点，(i) 首先，把充当谓语核心的动词性成分后面的体态算子跟这个动词性成分构成附加关系，(ii) 然后，依次向前，跟这个复杂的动词结构前面的各个时体算子或否定算子渐次构成偏正关系，并跟这个复杂的动词结构前面的各个模态算子渐次构成述宾关系。当然，这种句法树可能仍然是不完整的，因为在语义关系标注时，对于方式性成分没有作出标记。为了句法分析的完整，可以默认这些没有标记的方式性成分都是状语，依次跟其后的谓词性成分构成偏正关系。这样，这种句法树就基本完全了。为了节省篇幅，图解从略。

可见，通过基于论元结构的语义标注路线，不仅可以把语义关系标注可能带来的语义网库，有效地简化为一种扁平的依存树库；还可以从这种语义关系库中抽取出一棵棵句法树，从而把语义树库还原为句法树库，直接建立起语义结构和句法结构之间的映射关系。

5.2　语义关系标注语料库的作用

标注了上文所述的三种层面的语义关系的语料库，可以为信息

抽取提供强大的语义资源。比如,在一定程度上,论旨角色对应于信息抽取模板上的模板元素,论元之间的论旨角色关系对应于模板元素之间的事件关系;而否定、时体、模态语义可以对命题所表示的事件的类型和及其真实性作出约束,篇章关系又可以对代词和指示词的所指求解提供帮助。① 更进一步,这种语料库为机器学习和自动地识别句子各成分之间的语义关系,特别是对于机器学习和自动地发现篇章知识,提供十分精炼的训练语料。例如:

a. [邢云$_i$]$_{Ex}$,{1952年}$_T$ 生,[e_i]$_{Ex}$[V][大学文化]$_{Re}$。[e_i]$_{Ex}$历任[内蒙古伊克昭盟副盟长、盟委副书记、盟长]$_{Re}$,[e_i]$_{Ex}${1996年10月起}$_T$ 任[盟委书记、盟人大工委主任]$_{Re}$。

b. [罗多尔佛·塞尔特扎·塞维里诺]$_{Ex}${五日}$_T${在雅加达}$_L$ 正式就任[东南亚国家联盟(东盟)新一任秘书长]$_{Re}$。[塞维里诺$_i$]$_{Ex}${于去年七月}$_T${在吉隆坡召开的第三十届东盟外长会议$_j$上}$_L$ 被[e_j]$_A$ 任命为[东盟秘书长]$_{Re}$。[他$_i$ 的任期]$_{Th}$〈将〉$_{fut}${于二OO二年}$_T$ 结束。[塞维里诺$_i$]$_{Th}$是[菲律宾的一位外交家]$_{Re}$,[e_i]$_{Ex}$〈曾〉$_{past}$ 先后担任[菲律宾驻美国、中国和马来西亚等国的外交使节]$_{Re}$。{一九九二至一九九七年间}$_T$,[塞维里诺$_i$]$_{Ex}$任[菲外交部副部长]$_{Re}$,[e_i]$_{Ex}$ 负责[与东盟有关的事务等]$_{Re}$。

c. [阿尔巴尼亚前人民议会主席团主席列希]$_{Ex}${一日晚}$_T${在地拉那}$_L$ 病逝,[终年]$_{Th}$[V][八十五岁]$_{Ra}$。[列希$_i$]$_{Th}$是[阿尔巴尼亚反法西斯民族解放战争中的杰出人物]$_{Re}$。[列希i]Ex〈曾〉$_{past}$ 连续担任〈过〉$_{past-perf}$[十一届议员]$_{Re}$,[e_i]$_{Ex}${并}$_{ADD}${在一九五三年至一九八二年期间}$_T$ 担任[阿尔巴尼亚人民议会主席团主席国家元首之职]$_{Re}$。

d. [尼日尔警方]$_A${日前$_?$}$_T$ 破获[一个([e_i]$_A$ 企图暗杀[迈纳萨拉总统]$_P$ 的)阴谋团伙$_i$]$_P$。[尼日尔通讯社]$_A${二日}$_T$报道说,[[被捕的三名团伙成员]$_A$ 供认,[[他们$_i$]$_A$〈原〉$_{past}$ 定

① 详见袁毓林(2005a,b)。

[{于去年，十二月二十九日}$_{T}$ 行动]$_{Re}$，[暗杀对象]$_{Th}${除总统以外}$_{Re}$，还包括[几名政府重要成员]$_{Re}$]$_{Re}$]$_{Re}$。{据报道}$_{M}$，[该暗杀团伙$_{i}$的主谋]$_{Th}$是[哈马·阿马杜$_{j}$]$_{Re}$。[他$_{j}$]$_{Ex}$现任[尼(日尔)反对派"社会发展全国运动"总书记]$_{Re}$，[e_{j}]$_{Ex}${过去}$_{T}$〈曾〉$_{past}$担任〈过〉$_{pas\text{-}perf}$[政府总理]$_{Re}$。[警方]$_{A}${一日}$_{T}$〈已〉$_{past}$逮捕〈了〉$_{perf}$[哈马·阿马杜$_{j}$]$_{P}$，{但}$_{VER}$[他$_{j}$]$_{A}$〈不〉$_{neg}${承认[[e_{j}]$_{Th}$[跟这起暗杀活动]$_{Re}$有关]$_{Re}$}，[e_{j}]$_{A}$称[[自己$_{j}$]$_{P}$[被人]$_{A}$诬陷]$_{Re}$。

从篇章结构的角度看，生平介绍类文本通常用顺叙的写作手法，顺次交代主体论元的出生、学习、工作、任职经历等事项；这种顺叙往往要用一组顺序性的时间论元来显性地标志，如例 a 中的"1952 年……1986 年 10 月起……"。"就职、就任"等担任动词的后续句往往要交代获任的经过等情况，即整个话语采用倒叙的手法；其表层标志是使用逆序性的时间词语系列，如例 b 中的"五日……去年七月……"等。有时还要倒叙主体性论元在过去的任职经历，其表层标志是使用过去时标记"曾(经)"，然后是一组顺序性的时间词语，如本例中的"……曾……一九九二至一九九七年间……"。例 c 显示，讣告类文体通常要用倒叙的写作手法，先交代某人的死亡时间、地点等事项；然后交代该人的生平事迹。这种倒叙往往要用"曾"等表示过去的时体算子来显性地标志。本例在倒叙其任职经历时，还用了语篇关联词语"并"作标记。例 d 的篇章结构比较复杂，基本的叙述方式是顺叙(……日前破获……供认……〔……〕……一日已逮捕……)，但是其中又有方括号中的插叙(……主谋是……)，在插叙中因为涉及担任事件所以又引出倒叙(……现任……过去曾担任过……)。这种复杂的叙述安排，在表层结构上都有时间论元或时体算子作形式标志；并且，最后一句的主体论元转换时，还用了语篇关联词语"但"。

可见，高层次的顺序、倒叙、插叙等篇章构造方式，会在低层次的论元结构、逻辑结构和篇章关联词语上露出些许蛛丝马迹，认真地总结这种宏观的篇章结构的表层标记，将有助于机器自动地进行文本分类；当然，这对信息抽取时的事件模板类型的匹配，也将有约束作用。

6 结 语

上文给出了一种基于论元结构的汉语语义关系的标注体系：(i) 以谓词的论旨结构为基础，给谓词所支配的各个论元标注论旨角色；(ii) 给附加在论元结构上的否定、时体和模态算子等逻辑成分标注语义功能及其辖域，给指代词标注照应关系；(iii) 给联结不同的论元结构的语篇衔接词语标注篇章功能及其配对关系。为这三个层面上的各种语义关系设计了便于记忆的标记，形成了一套可扩充的标记集(tag set)。并且，为各种语义关系的标注制定了比较具体的操作规范，更为具体的细则(specification)有待于在更大规模的语料标注实践中逐步形成。最后，展示怎样从经过上述语义标注的语料上自动地为句子建立语义依存树和句法关系树，还讨论了这种语料库在信息抽取、机器学习等领域的运用。通过对数万字新闻文本的手工标注，显示出这套标注体系对真实语料具有较好的适应性和较高的语义关系信息的覆盖率。

参考文献

陈　平 (1991)《现代语言学研究——理论、方法与事实》。重庆：重庆出版社。

廖秋忠 (1992)《廖秋忠文集》，北京：北京语言学院出版社。

徐烈炯 (1988)《生成语法理论》，上海：外语教学与研究出版社。

袁毓林 (2002a)《信息抽取的语义知识资源研究》，《中文信息学报》第 5 期。

袁毓林 (2002b)《论元角色的层级关系和语义特征》，《世界汉语教学》第 3 期。

袁毓林 (2003)《走向多层面互动的汉语研究》，《语言科学》第 6 期。

袁毓林 (2005a)《用动词的论元结构跟事件模板相匹配》，《中文信息学报》第 5 期。

袁毓林 (2005b)《用逻辑和篇章知识来约束模板匹配》，《中文信息学报》第 4 期。

Brown, Gillian & Yule, George (1983) *Discourse Analysis*. Cambridge: Cambridge University Press.

de Beaugrande, Robert-Alain & Dressler, Wolfgang Ulrich (1981) *Introduction*

to Text Linguistics. London: Longman.

Halliday, M. A. K. & Hasan, R. (1976) *Cohesion in English*. London: Longman.

Shi, Dingxu (2000) Topic and Topic-comment Constructions in Mandarin Chinese. *Language* 76: 383—408.

van Dijk, Teun A. (1985) *Handbook of Discourse*, Vol. 2: *Dimensions of Discourse*. London: Academic Press.

2004 年 8 月初稿，2004 年 9 月改定

附录

汉语语义关系网库标记集

0 引　言

这套汉语语义关系网库标记集包括三个子集：论旨角色标记集、逻辑关系标记集、语篇关系标记集，基本覆盖谓词性成分的论元结构、附加在论元结构上的逻辑语义结构、联结论元结构的话语篇章结构三个层面上的主要的语义关系。这套标记集共有 52 个(54 个减去 2 个重复的)标记。

1 论旨角色标记集

序号	标记代码	标记名称	帮助记忆的说明	举　例
1	A	施事	agent 的首字母	弟弟吃冰棍
2	Se	感事	sentient 的前两个字母	刘芳认识局长
3	Ex	经事	experiencer 的前两个字母	王平担任校长
4	Cau	致事	causer 的前三个字母	忧愁使人苍老
5	Th	主事	theme 的前两个字母	积雪融化了
6	P	受事	patient 的首字母	厂长批评小陈
7	D	与事	dative 的首字母	送爷爷一瓶酒
8	R	结果	result 的首字母	盖一座小楼
9	Ta	对象	target 的前两个字母	喜欢山水画
10	Re	系事	relative 的前两个字母	产权属于学校
11	I	工具	instrument 的首字母	用锤子砸
12	Ma	材料	material 的前两个字母	用毛线织
13	M	方式	manner 的首字母	用花腔唱
14	Rn	原因	reason 的首尾字母	因暴雨停飞
15	Ai	目的	aim 的前两个字母	为和平奋斗
16	T	时间	time 的首字母	在晚上出门
17	L	处所	location 的首字母	在家里休息
18	So	源点	source 的前两个字母	从上海出发

序号	标记代码	标记名称	帮助记忆的说明	举　　例
19	Go	终点	goal 的前两个字母	向北京推进
20	Pa	路径	path 的前两个字母	经天津回国
21	Ra	范围	range 的前两个字母	僵持了半年
22	TOP	话题	topic 的前三个字母	这人嘴太快
23	COM	说明	comment 的前三个字母	那菜味道辣

说明：(1) 论旨角色的标记代码，采用首字母大写的方式。(2) 论旨角色的标记代码，尽可能采用语言学文献上通用的缩写形式，以便记忆和人工标注及校对。(3) 话题和说明本来是话语、篇章结构层面上的功能性成分，因此一律采用大写字母。(4) 在我们的标注体系中，只有当话题性成分跟后面说明性成分中的谓词没有论旨角色关系时，才标注这种话语功能。也就是说，只有当体词性成分没有论旨角色可标注时，才用话语功能角色济其穷；因此，这里把话题、说明暂且放在论旨角色标记集中。

2 逻辑关系标记集

序号	标记代码	标记名称	帮助记忆的说明	例　　词
1	neg	否定算子	negator 的前三个字母	不、没有、未
2	mod	模态算子	modality 的前三个字母	能、可以、该
3	fut	将来时	future 的前三个字母	将、即将、快
4	past	过去时	整个英语单词 past	刚、已、曾经
5	prog	进行体	progressive 前四个字母	着、呢、正在
6	perf	完成体	perfect 的前四个字母	了
7	pres-perf	现在完成体	present 的前四个字母	来着、来的
8	past-perf	过去完成体	过去完成的英语缩写	过

说明：(1) 逻辑语义关系的标记代码，一律采用小写字母。(2) 逻辑语义关系的标记代码，尽可能采用语言学文献上通用的缩写形式，以便记忆和人工标注及校对。

3　语篇关系标记集

序号	标记代码	标记名称	帮助记忆的说明	例　词
1	COR	并列关系	coordinate 的前几个字母	既…又…、…也…
2	COR-i	并列选枝	同一下标表示同一关系	既…、又…、也…
3	ADD	递进关系	additive 的前三个字母	不但…而且…
4	ADD-i	递进选枝	同一下标表示同一关系	不但…、而且…
5	ALT	选择关系	alternative 的前三个字母	或者…或者…
6	ALT-i	选择选枝	同一下标表示同一关系	或者…、或者…
7	TEM	连贯关系	temporal 的前三个字母	首先…然后…
8	TEM-i	连贯选枝	同一下标表示同一关系	首先…、然后…
9	CON	条件关系	conditional 的前三个字母	只要…就…
10	CON-i	条件选枝	同一下标表示同一关系	只有…、不管…
11	CSQ-i	结果选枝	consequence 音节首字母	才…、就…
12	CAS	因果关系	causal 的前几个字母	因为…所以…
13	CAS-i	原因选枝	cause 的前几个字母	由于…、既然…
14	CSQ-i	结果选枝	同一下标表示同一关系	因而…、那么…
15	SUP	假设关系	suppositional 前三个字母	如果…那么…
16	SUP-i	假设选枝	supposed 的前三个字母	假使…、即使…
17	CSQ-i	结果选枝	同一下标表示同一关系	那么…、否则…
18	VER	转折关系	adversative 的中间字母	虽然…但是…
19	CES-i	让步选枝	concessive 的中间字母	虽然…、诚然…
20	VER-i	转折选枝	同一下标表示同一关系	然而…、不过…
21	PUR	目的关系	purposive 的前三个字母	…以便…
22	PUR-i	目的选枝	同一下标表示同一关系	为了…、省得…
23	MEN-i	手段选枝	means 的前三个字母	没有专用的连词

说明：(1) 篇章语义关系的标记代码，一律采用大写字母。(2) 篇章语义关系的标记代码，尽可能采用语言学文献上通用的缩写，以便记忆和人工标注及校对。(3) 并列、递进、选择和连贯这四种承接性的关系，其中的选枝用相同的标记，并暗示其选枝可以是两项以上的。(4) 条件、因果、假使、转折和目的这五种条件性的关系，其中的

选枝基本上是条件与结果的两项对举，因此用不同的标记。（5）选枝标记中的下标 i 是一个代数值，代表 i，j，k…以降的各个字母。相同的字母表示处于同一种条件—结果关系之中。

新闻语体真实文本的语义标注的实践

本文以信息抽取等语言信息处理工程为应用背景，根据袁毓林(2004)提出的“基于论元结构的语义标注的体系和规范”，选择新闻报道中关于职务调动的真实文本，分别从论元结构、逻辑结构和篇章结构三个方面，进行语义关系标注的实践。这些文本分为段落、单条简讯和全文三种类型，藉此可以发现不同长度单位的文本在语义结构和语义表达方面的若干差异。

0 引 言

本文以信息抽取等语言信息处理工程为应用背景，分别选择新闻报道中关于职务调动的真实文本进行语义标注的实践。关于语义标注的理论和规约，我们有专文讨论和说明，在这里也随文作出说明和交代。这些文本分为段落、单条简讯和全文三种类型，下面依次列出。

1 新闻段落的语义标注

这些文本都是新闻报道中的比较完整的段落，主题都是关于职务变动的。来自孙斌(2000)，是他的信息抽取模型(InfoX)的应用实例“职务变动”的测试语料。这些文本的编号原文没有，是我们为了查找和核对的方便而加上去的。

(1) [本报]$_{A}${伊斯兰堡}$_{L}${12月31日}$_{T}$电[记者王南]$_{A}$报道：[[巴基斯坦穆斯林联盟谢里夫派候选人、原最高法院大法官穆罕默德·拉斐克·塔拉尔]$_{Ex}$，{今天}$_{T}${在巴国民议会、参议院以及各省议会选举中}$_{L}$，当选[巴基斯坦第九任总统]$_{Re}$，

[任期]$_{Th}$[V][5 年]$_{Re}$。[塔拉尔]$_{Ex/A}$〈将〉$_{fut}${于明天}$_{T}$ 宣誓就职]$_{Re}$。

说明：(i) 新闻报道的引语中的"电"可以看作是个名词，这个单音节名词还没有收入《现代汉语词典》。其相应的双音节名词是"电讯"，意思是：通过电话、电报或其他无线电设备传播的消息。这个双音名词已收入《现代汉语词典》。为了标注的简单和语义角色关系的清晰，我们暂时忽略其中的名词化之后的指称性意义(designative meaning)，而径直标注其名词化之前的谓词性结构的陈述性意义关系(assertive relation)。即假定"用无线电传播的消息"这种名词性结构是从"用无线电传播消息"这种动词性结构上，通过名词化转换而派生出来的。也就是说，姑且把这里的"电"解释为是动词，意思是：通过电话、电报或其他无线电设备传播消息。参见第 3 部分第(1)例的说明中对"专电"的解释。"本报"之类表示机构的名词性成分，在这种场合具有拟人化的语用特点，因而具有[＋human]的语义特征，充当动词"电"的施事(agent，简写为 A)。

(ii) 新闻报道的引语中的处所(location，简写为 L)和时间(time，简写为 T)论元通常不用介词"在、于"等来引导。

(iii) "报道"类动词通常带命题性超级论元，这种论元的论旨角色是系事(relative，简写为 Re)。这种超级论元中一定包含动词性成分，这种动词性成分的论元结构也需要进行语义标注。可以通过括号来表示和辨识其嵌套层次。根据同样的道理，当我们把"电"处理为动词之后，"电"所支配的后续一连串句子也是它的系事。为了简便，我们在这里作一个总的说明，然后在实际作语义标注时省去了这一层次。

(iv) "当选"类任职动词的主体性论元的论旨角色可以归入经事(experiencer，简写为 Ex)，即经历了某种变化的具有感知性的主体；其客体性论元的论旨角色可以归入系事(relative，简写为 Re)，即在事件结构中跟主体性论元相关的事物，比如经事所担任的某种职务。动词"就职"一般不单独使用，通常跟"宣誓"连用。"宣誓就职"的主体论元，具有一定的施动性(causation)，可以归入施事。但

是，为了跟其他任职动词的论元结构相协调，把"宣誓就职"的主体论元也可以归入经事。当然，为了周全，可以把这两种论旨角色都标注上去；并且，在这两种标记之间用斜撇号(/)隔开。

(v) 像"任期"等既无施动性、又无感知性的主体论元，可以归入主事(theme，简写为 Th)。

(vi) 为了区别，动词性成分的必有论元用方括号标志，非必有论元用花括号标志。

(vii) "任期 5 年"可以看作是"任期为 5 年"的省略形式，其中，所省略的动词标记为[V]。参看第 2 个文本中的最后一句话。这里"5 年"是真宾语，其论旨角色可以归入系事。

(viii) 副词"将"表示将来时(future tense，简写为 fut)。

(ix) 从篇章结构的角度看，这一段中有三个句子，每个句子都有一个时间论元"12 月 31 日……今天……明天……"正好顺序把句子组织成一段话语，其中引语中的所指明确的"12 月 31 日"跟正文中的"今天、明天"构成参照语与指示语之间的照应和索引关系；再加上正文中的谓语核心动词"当选……宣誓就职……"之间的顺序性的先后事件关系，正好显示出整个段落是按照时间顺序、用顺叙这种叙述手法来写作的。

(2) {据[新华社]$_{A}${香港}$_{L}${1 月 3 日}$_{T}$ 电}$_{M}$[香港特区政府]$_{A}${昨天}$_{T}$ 公布〈了〉$_{perf}$[香港特区基本法推广督导委员会成员名单]$_{P}$，[政务司长陈方安生$_{i}$]$_{Ex}$出任[委员会主席]$_{Re}$，[高苕华$_{j}$]$_{Ex}$任[副主席]$_{Re}$，[其他成员$_{k}$]$_{Th}$包括[8 位特区政府高级官员和 12 位社会不同界别人士]$_{Re}$。[他们$_{i+j+k}$的任期]$_{Th}$均为[两年]$_{Re}$。

说明：(i) 介词"据"可以带动词性成分作宾语，比如："据报道、据估计、据小王说"。这种介词的宾语的论旨角色可以归入系事，但我们认为介词是论旨角色的标记，所以不必标注介词宾语相对于介词的论旨角色，而是标注整个介宾结构相对于后面的谓语核心动词的论旨角色。

(ii) 这种介词结构可以独立作报道的引语。其论旨角色可以归

入方式(manner,简写为M)。支配这种论元角色的谓词及其必有论元是“本报记者某某某报道”一类结构(比如,第1条文本中的“记者王南报道”),可以省略。一旦省略了这种结构,便使“据……电”一类结构失去了依托,其论元角色也模糊了。

(iii) 助词“了”表示完成体(perfective aspect,简写为perf)。

(iv) 后句中的代词“他们”跟前句中的先行词“陈方安生”、“高苕华”和“其他成员”之间的照应关系,正好起到把句子衔接成为语篇的连贯(cohesion)作用。

(3) {附记}$_{M}$:{目前}$_{T}$,[徐虎$_{i}$]$_{A}$ 既做[老师]$_{Re}$又当[学生]$_{Re}$,[e_{i}]$_{A}$ 奔波于{两个课堂}$_{L}$。[上海市房屋土地管理局$_{j}$]$_{A}$ {为广泛推广徐虎精神,提高居民住宅管理水平}$_{Ai}$,成立〈了〉$_{perf}$ [一所“徐虎学校$_{k}$”]$_{R}$。[上海市房管系统35岁以下的青工]$_{Ex}$ 都{在这里$_{j}$}$_{L}$ 脱产培训。[徐虎]$_{Ex}$被[e_{j}]$_{A}$任命为[这所学校$_{k}$的校长]$_{Re}$。

说明:(i)“附记”类篇章说明词语的论旨角色可以标记为方式,详见第2条文本中的说明(ii)。“既……又……”等小句内部的关联词语不作标记。“都”等副词、“脱产”等表示方式的状语都不加标记。

(ii) 后续句中承前省略的论元成分是空语类(empty category),用[e]标记;并用下标来表示它跟其先行成分的语义同指关系。空语类通常是可以明确地补出来,并且补出来的语句形式一般是合语法的,只有少数句子因为特定句法结构的限制而不合语法。

(iii) 介词是论元角色的标志,俗称“格标记”(case marker)。为了醒目,跟动词一样,用着重点标注。另外约定:在动词之前,整个介词结构都置于一个括号中,即作为一个论元成分;但是在动词之后置于括号之外,即把动词和其后的介词看作一个动词性成分。① 当“被”等引导必有论元的介词之后不出现宾语时,用空语类[e]作标记,并加上同指下标和论旨角色标记。这种带空语类作宾语的介词,

① 这样处理有句法、语义和音系学上的考虑,详见袁毓林(2003)及其所引的参考文献。

因为作为动词短语的一部分，所以不用放在方括号中，即单独把空语类放在括号中。

(iv) 目的(aim，简写为 Ai)是非必有论元，置于花括号中。因为我们约定：谓词性成分的必有论元用方括号套起来，可有论元用花括号套起来，一律在括号后加上论旨角色标记。

(v) 代词性成分和其先行成分的语义同指关系，通过加共同的下标来标注。

(vi) 后句中的指示词"这里"跟前句中的先行参照词"徐虎学校"之间的照应关系，正好起到衔接语篇、呼应前后的作用。

(4) [刘沈明$_i$]$_{Ex}$〈原〉$_{past}${在福建省海洋渔业公司}$_L$ 当[车间主任]$_{Re}$，[e$_i$]$_{Ex}$下岗{5 年多}$_{Ra}$，[他]$_A$ 到处打工。{去年十月底}$_T$，{当他重新拿起笔来准备应试的时候}$_T$，[那种久违了的"找到组织"的感觉]$_{Th}$ 重又回到[身上]$_{L(G)}$。

说明：(i) "重、又"等重复副词暂时不作语义分析，因而也不加语义标记。

(ii) 动词性成分之后的"5 年多"等时量成分是一种准宾语，在论旨角色上可以归入范围(range，简写为 Ra)。

(iii) "身上"等处所性成分是"回到"等表示移动的动词性成分的必有论元，在论旨角色上可以归入表示终点或目标(goal)的处所(简写为 L(G))。像处所这类论旨角色，对于有的动词性成分来说是必有论元(如本例所示)，对于有的动词性成分来说却不是必有论元。

(iv) 从上例可以发现，同一个动词可以带一个以上的时间论元，这些论元的排列顺序是从较大的时间到较小的时间。

(v) 时间副词"原"表示过去时(past tense，简写为 past)。

(vi) 指示词"这、那"直接修饰、限定其先行语构成的复指性结构时，为了简单，可以不标注其照应关系。

(5a){1996 年初}$_T$，[李长水$_i$]$_{Ex}$ 担任〈了〉$_{perf}$[市公安局长、党委书记]$_{Re}$，[e$_i$]$_{Ex}$ 负责[市公安局的全面工作]$_{Re}$。{[他$_i$]$_{Ex}$ 上任后}$_T$，[PRO$_j$]$_A$ 立即召开[党委会]$_R$，[PRO$_j$]$_A$ 研究[(跟)群众息息相关的治安问题]$_P$。[他$_i$]$_A$ 提出，[{在社会治安综合治

理中}$_L$，[公安系统$_k$的责任]$_{Th}$最大，[e_k]$_{Th}$要[[e_k]$_A$[把工作重心]$_P$放在[加强管理和防范上来]$_{L(G)}$，[e_k]$_A$依靠和发动[群众]$_P$，[e_k]$_A$走[“联户联防”的治安路子]$_{Ma}$]$_{Re}$]$_{Re}$。

说明：(i) PRO代表隐含在语境中的必有性论元成分，由于句法结构上的限制，一般是不可补出来的。比如，上例中“召开党委会”的应该是“李常水”和“党委一班人”。

(ii) 圆括号中的介词“跟”原文没有，是我们根据文意补上去的。

(iii) 助动词“要”等的主体性论元(主语)在论旨角色上可以归入主事，客体性论元(宾语)在论旨角色上可以归入系事。另一种办法是把助动词看作是一种模态算子，可以作如下这种标记：

(5b){1996年初}$_T$，[李长水$_i$]$_{Ex}$担任〈了〉$_{perf}$[市公安局长、党委书记]$_{Re}$，[e_i]$_{Ex}$负责[市公安局的全面工作]$_{Re}$。{[他$_i$]$_{Ex}$上任后}$_T$，[PRO$_j$]$_A$立即召开[党委会]$_R$，[PRO$_j$]$_A$研究[(跟)群众息息相关的治安问题]$_P$。[他$_i$]$_A$提出，[{在社会治安综合治理中}$_L$，[公安系统$_k$的责任]$_{Th}$最大，[e_k]$_A$〈要〉$_{mod}${[把工作重心]$_P$放在[加强管理和防范上来]$_{L(G)}$，依靠和发动[群众]$_P$，走[“联户联防”的治安路子]$_{Ma}$}]$_{Re}$。

说明：(i) 更为简单和可靠的办法是：把助动词看作是一种附加在其所支配的动词的论元结构之上的模态算子，表示某种情态(modality，简写为mod)，并用花括号标志其辖域。这样，不仅标明了其所支配的动词的论元结构关系，而且还标明了句子在情态上表示的是一种非现实的断言(irrealis assertion)。下面原则上都按这种方式标注。

(ii) 后两句中的人称词“他”跟前句中的先行词“李长水”之间的照应关系，正好起到连贯语篇、呼应前后的作用。

(6) {1984年}$_T$，[广昌]$_A$成立〈了〉$_{perf}$[全国第一个白莲科研所]$_R$，[刘光亮]$_{Ex}$担任[所长]$_{Re}$，[e_i]$_{Se}$倍感[肩上担子的沉重]$_{Re}$。

说明：(i) 从上例可以看出，空语类的论旨角色可以跟其先行语

不一样。

(ii)“倍感”等心理感觉类动词的主体论元是感事(sentient,简写为 Se),表示感觉内容的宾语在论旨角色上可以归入系事。

(iii) 从篇章结构的角度看,前后小句中的谓语核心动词“成立……担任……”之间的顺序性的先后事件关系,正好显示出整个句子是按照时间顺序、用顺叙这种叙述手法来写作的。

(7) {1954 年}$_{T}$,[24 岁的梁惠珍$_{i}$]$_{Ex}$ 开始〈了〉$_{perf}$[行医生涯]$_{Re}$。[她$_{i}$]$_{A}${先}$_{TEM\text{-}i}$是{在一个县卫生所}$_{L}$ 工作,[e$_{i}$]$_{Ex}${后来}$_{T\&TEM\text{-}i}$成为[海南省屯昌县人民医院的妇产科主任]$_{Re}$。[长期的临床实践]$_{Cau}$使[她]$_{P\&Se}$看到[许多不育妇女的痛苦]$_{Re}$,[梁惠珍$_{i}$]$_{Se}$决心[[e$_{i}$]$_{A}${用更好的方法}$_{I}$解决[这个问题]$_{P}$]$_{Re}$。

说明:(i)“使”等使令动词的主体论元可以归入致事(causer,简写为 Cau),其客体论元可以归入受事。同时,使令动词一定要有后续动词,并且其受事正好是后续动词的主体性论元,这就是传统语法所谓的“兼语”。我们通过合取符号 & 来在一个论元上同时标记其相对于前后动词性成分的两种论旨角色。这从语法理论上讲,会违反生成语法的 GB 理论中的论旨原则:一个论元只能担任一种论旨角色,一种论旨角色只能赋予一个论元。[①] 但是,从语言信息处理工程的角度上讲,这种合成标记法比较经济,也便于识别和处理,具有更高的效率。这种具有客体和主体双重论旨角色特征的成分也可以简单地归入经事,即某种事件的经历者。

(ii) 介词“用”通常引进工具论元(instrument,简写为 I)。

(iii) 为了显示语篇结构关系,我们把副词“先”等能把小句、句子、甚至段落等较大的语言单位连接起来、使之成为句子、话语、甚至篇章的词语或其他表层结构上的词汇特征,笼统地叫做话语衔接词语(cohesive expression,简写为 COH)。这里的“先、后来”表示时间性的连贯关系(temporal relation,简写为 TEM),我们用相同的下标来表示一组关联词语处于相同的衔接关系之中。

① 关于论旨原则,详见徐烈炯(1988)第 271 页。

(iv) 后句中的两个代词“她”跟前句中的先行词“梁惠珍”之间的照应关系，正好起到语篇衔接作用。最后一个小句的主体论元不用代词“她”、而是直接用指称词语“梁惠珍”，可能是为了避免行文措辞的重复单调。即用了修辞上“避复”的修辞手法。

(8) {1989 年}$_T$[[梁惠珍$_i$]$_{Ex}$退休后]$_T$，[[e_i]与丈夫$_j$]$_A$ 回到[老家湛江]$_{L(G)}$，[e_j]$_A$ 开办〈了〉$_{perf}$[“惠珍联合医院$_k$”]$_R$，[e_k]专治[男女不育症]$_P$。

说明：(i) “开办”的施事是空语类，其先行语显然是“[e_i]与丈夫”。而这个先行语中的空语类 e_i 的先行语是其先行句中的“梁惠珍”。

(ii) 上面(7)(8)两个文本应该是一个段落，前后句子中的衔接词语“先”和时间论元“后来……梁惠珍退休后……”把句子组织成一段话语，并显示出整个段落是按照时间顺序、用顺叙这种叙述手法来写作的。

(9) [“黎明俱乐部”]$_{Th}$共有[18 名成员]$_{Re}$。{在成立大会上}$_L$，[白滨一良$_i$]$_{Ex}$被[$e_?$]$_A$ 选为[代表]$_{Re}$，[木庭健太郎$_j$]$_{Ex}$被[$e_?$]$_A$ 选为[干事长]$_{Re}$。[他们$_{i+j}$]$_{Ex}$都〈曾〉$_{past}$担任[原公明党副书记长]$_{Re}$。

说明：(i) 我们用加号来连接复数性代词后面的几个同指下标。

(ii) 当“被”等表示被动的介词之后不出现宾语(通常是施事)时，把“被 VP”看作是一个动词性成分。参看第(11)条，说明(i)。当然也可以把这个省去的主体性论元用空语类的形式补充出来，当其先行词不明确时，可以用问号作下标。

(iii) 后句中的代词“他们”跟前句中的先行词“白滨一良”和“木庭键太郎”之间的照应关系，正好起到衔接语篇、连贯语义的作用。

(10) [捷克总统哈维尔]$_A${二日}$_T${在布拉格宫}$_L$ 任命〈了〉$_{perf}$[捷克新政府成员]$_{Ex/Re}$。[([由托绍夫斯基]$_{Ex}$ 出任[总理]$_{Re}$的)新政府$_i$]$_{Th}$共有[十八名成员]$_{Re}$，{其中$_i$}$_L$[新入围的人数]$_{Th}$占[一半]$_{Re}$。[本届政府中]$_L$ 有[七名无党派人士]$_{Re}$，

[这]$_{Th}$是[历届政府中少有的现象]$_{Re}$。[以克劳斯为首的捷克原政府]$_{Ex}$是{一九九七年十一月三十日}$_{T}$被迫辞职的。[托绍夫斯基]$_{A}${二日}$_{T}${在新政府成立后会见记者时}$_{T}$表示,[[本届政府$_{j}$优先考虑的问题]$_{Th}$是,[[e_{j}]$_{A}$积极争取加入[北约和欧盟]$_{L(G)}$,[e_{j}]$_{A}$[同经济犯罪行为]$_{Re}$作斗争,[e_{j}]$_{A}$加快[经济改革]$_{P}$]$_{Re}$]$_{Re}$。

说明:(i) 当"任命"等动词的客体论元是指人名词时,其论旨角色可以归入经事;当"任命"等动词的客体论元是职务名词时,其论旨角色可以归入系事。上例中的"捷克新政府成员"似乎既涉及人员,又涉及职务;为了周全,特意把这两种论旨角色都标记上去,并用斜撇号表示这两种角色之间是一种析取关系。

(ii) 我们把"争取加入"、"作斗争"等动词性结构当作一个谓词整体,不加分析。

(iii)"是……的"等强调性成分、"被迫"等方式副词暂不标记。

(iv) 从篇章结构的角度看,这一段整体上是用顺叙的手法来组织句子的,时间论元和动词"……二日……任命……出任……,〔一九九七年十一月三十日……辞职……,〕……二日……成立……"之间的顺序性的先后事件关系,正好显示了这一点。但是,中间用插叙的手法交代了原政府辞职这一事件(用方括号标示)。这种插叙是通过时间论元的转变来显示的,具体地说是把时间论元"12 月 31 日"插在两个同指的时间论元"二日"之间,来显示要插入一段叙述。可见,高层次的写作叙述手段,也会在低层次的论元结构中落下些许蛛丝马迹。

(11) [塔拉尔$_{i}$]$_{Ex}${于1997年$_{j}$3月}$_{T}$当选[巴参议院议员]$_{Re}$,[e_{i}]$_{Ex}${同年$_{j}$12月15日}$_{T}$[被执政党穆斯林联盟谢派]$_{A}$提名为[总统候选人]$_{Re}$。

说明:(i) 当"被"等表示被动的介词之后出现宾语(通常是施事)时,把"被 NP"看作是一个论元成分。参看第(9)条,说明(ii)。

(ii)"同年"等指示性词语(deixis),标明其参照性的先行词语。"同年"的语义跟其参照性词语的所指相同。

(iii) 后一小句中的指示词“同年”跟前一小句中的参照性的先行词“1997 年”之间的照应关系，正好起到把两个小句衔接成一个大句的作用。

(12) [恩佐]$_A$ 宣布，[[南非总统曼德拉]$_A$〈已〉$_{past}$ 任命[南非驻中国研究中心主任戴克瑞]$_{Ex}$ 为[首任驻华大使]$_{Re}$]$_{Re}$。

说明：(i) 为了一致和醒目，动词之后跟动词不连续的介词也放在表示论元成分的方括号之外；即把“任命……为”看作是一个动词性结构，属于一种不连续的动词性成分。参看第(3)条，说明(ii)。

(13) {据[新华社]$_A${伊斯兰堡$_i$}$_L${1 月 1 日$_j$}$_T$ 电]}$_M$([记者杨士龙]$_A$[V])[新当选的巴基斯坦总统拉菲克·塔拉尔$_k$]$_A${今天$_j$}$_T${在这里$_i$}$_L$ 宣誓就职。[他$_k$]$_{Th}$ 是[巴建国以来的第九任总统]$_{Re}$。

说明：(i) 在新闻引语中，记者姓名之后，经常省略“报道”一类动词性成分。参看第 1、20 个文本。

(ii) “今天、这里”等多个指示词语，分别用不同的下标来标明其跟各自的参照性词语的照应关系。

(iii) 后一句中的人称词“他”跟前一句中的先行词“拉菲克·塔拉尔”之间的照应关系，正好起到连贯语篇、呼应前后的作用。

(14) [越南国防部]$_A${三日}$_T$ 举行〈了〉$_{perf}$[对一批高级将领的授衔、授职仪式]$_R$。{根据越南国家主席和政府总理的决定}$_M$，[国防部长范文茶]$_{Ex}$[由中将]$_{Re1/L(S)}$ 提升为[上将]$_{Re2/L(G)}$。[陶仲历少将]$_{Ex}$ 任[国防部副部长、越南人民军总参谋长]$_{Re}$，[范清银中将]$_{Ex}$ 任[人民军总政治局主任]$_{Re}$，[阮华盛少将]$_{Ex}$ 任[人民军总技术局主任]$_{Re}$，[阮文沱大校]$_{Ex}$ 任[人民军总后勤局主任]$_{Re}$。

说明：(i) “举行”等制作性动词的客体论元的论旨角色可以归入结果(result，简写为 R)。

(ii) 为了简单，可以把由“根据”引导的表示依据、前提的论元的论旨角色归入方式(manner，简写为 M)。

(iii) 在“由中级提升为上将”中,“中将”的论旨角色可以归入表示来源(source)的处所(简写为 L(S)),“上将”的论旨角色可以归入表示目标、终点的处所 L(G)。这里为了跟其他相关的职务变更动词的客体论元相协调一致,同时也分别把它们标记为 Re1 和 Re2。

(iv) 从篇章结构的角度看,这一段第一个句子(话题句)用的是机构名词作主体性论元,后面的句子(包括小句)都用从属于这一机构的个体名词作主体论元,这种不同的主体性论元在语义上的所属关系;再加上话题句中具有概括性的动词短语“举行……授衔、授职仪式”跟后续句中比较具体的动词“提升、任”之间在语义上的上下位关系,正好显示出整个段落是按照“先总后分”这种篇章组织方法来展开叙述的。

(15) [阿尔巴尼亚前人民议会主席团主席列希]$_{Ex}${一日晚}$_{T}${在地拉那}$_{L}$ 病逝,[终年]$_{Th}$[V][八十五岁]$_{Ra}$。[列希]$_{Th}$是[阿尔巴尼亚反法西斯民族解放战争中的杰出人物]$_{Re}$。[列希$_{i}$]$_{Ex}$〈曾〉$_{past}$ 连续担任〈过〉$_{past\text{-}perf}$[十一届议员]$_{Re}$,[e_i]$_{Ex}${并}$_{ADD}${在一九五三年至一九八二年期间}$_{T}$ 担任[阿尔巴尼亚人民议会主席团主席国家元首之职]$_{Re}$。

说明:(i)“终年”在表示“人去世时的年龄”时,只能作主语,并且只能以“数词+岁”作谓语。为了方便和系统,可以假定中间隐含了“是、为”一类谓词,标记为[V]。这样,从论旨角色上看,这里的“终年”是主事,“八十五岁”是范围。

(ii) 助词“过”表示经历过某种事件,可以归入过去完成体(past perfect,简写为 past-perf)。当“并”起联结小句的作用时,是语篇关联词语,表示递进关系(additive relation,简写为 ADD)。

(iii) 从篇章结构的角度看,讣告类文体通常要用到倒叙的写作手法,即先交代某人的死亡时间、地点等事项;然后交代该人的生平事迹。这种倒叙往往要用“曾”等表示过去的时体算子来显性地标志:本例在倒叙其任职经历时,还用了语篇关联词语“并”作标记。

(16) [龚德俊$_{i}$]$_{Ex}$ 是[北京中诚信租赁有限公司的董事长]$_{Re}$,[e_i]$_{Ex}$〈曾〉$_{past}${在中汽专用汽车珠海制造有限公司}$_{L}$ 任

〈过〉$_{past\text{-}perf}$[总经理]$_{Re}$。[他$_i$]$_A${先}$_{TEM\text{-}i}${从日本}$_{L(S)}$进口[原装马自达323$_j$]$_P$至[香港]$_{L(G)}$，[e$_i$]$_A$就地拆散[e$_j$]$_P$，{按配件}$_M$报关进口。{然后}$_{TEM\text{-}i}$，{再}$_{TEM\text{-}i}$[由广东、广西、湖北、四川等地的汽车修配厂]$_A$[将它们?]$_P$组装成[车]$_R$。{为掩人耳目}$_{Go}$，{给走私非法组装车}$_{Re}$找[一个合法的“户口”]$_P$，[他$_i$]$_A${又}$_{TEM\text{-}i}${以每辆车4000元的代价}$_M$，{从海南汽车工业公司北海公司}$_{L(S)}$购买〈了〉$_{perf}$[HMC5010的铭牌、标牌、合格证等一整套手续]$_P$。就是{凭着这貌似合法的伪装和一些不为人知的手段}$_M$，[龚德俊的这批“海马”]$_A$偷逃〈了〉$_{perf}$[关税、增值税、消费税]$_P$，畅行无阻地开进〈了〉$_{perf}$[京城]$_{L(G)}$。

说明：(i)“从日本”在论旨角色上可以归入表示来源的处所L(S)。

(ii)“为掩人耳目”的论旨角色可以归入目的(aim，简写为Ai)。

(iii)对于事件结构中的非主动的相关参与者，如果是有生性的，那么其论旨角色可以归入与事(dative，简写为D)；如果是非有生性的事物或事件，那么其论旨角色可以归入系事。这里的“给走私非法组装车”，其论旨角色可以归入系事。

(iv)这里的代词“它们”的先行语没有以显性的词汇形式出现，而是隐含在上下文语境中。其所指是：被拆散成为零部件的那批原装马自达323汽车。

(v)“龚德俊的这批“海马””通过拟人化的语用手段，充当施事角色。在“开进了京城”中，终点性处所“京城”是“开进”的必有论元。

(vi)“就是”等强调性成分、“畅行无阻”等方式状语都不加语义关系方面的标记。

(vii)语篇衔接词语“……先……然后，再……。……又……”等显示整个段落是按照顺叙的手法来组织篇章结构的。

(17){([陈云峰$_i$]$_A$读[电大]$_{Re/M}$)期间}$_T$，[他$_i$]$_{Se}$认识〈了〉$_{perf}${现在的妻子黄赛红$_j$}$_{Ta}$。[这位与他$_i$有相同经历的农村青年$_j$]$_A$，{从代课教书}$_{L(S)}${到做临时工}$_{L(G)}$，{处处}$_L$体现〈了〉$_{perf}$[([她$_j$]$_A$〈不〉$_{neg}${安于[生活现状]$_{Re}$}的)个性和坚强的毅

力]$_{P}$。{1994 年}$_{T}$，[黄赛红$_{j}$]$_{A/Ex/Se}$ 考取〈了〉$_{perf}$[浙江省政法管理干部学院]$_{P/Re/Ta}$，[e_j]$_{Ex}${通过三年的自费脱产学习}$_{M}$，{于1996 年 7 月$_{k}$}$_{T}$ 毕业。{也}$_{COR}$就{在这个月$_{k}$}$_{T}$，[陈云峰$_{i}$]$_{Ex}${因出色的工作成绩}$_{Rn}${从城关镇内设机构青马办事处}$_{L(S)}$调到[沙善办事处]$_{L(G)}$[e_i]$_{Ex}$当[副主任]$_{Re}$。

说明：(i) "认识"等心理感觉类动词的主体论元是感事 Se，表示感觉对象的宾语在论旨角色上可以归入对象(target，简写为 Ta)。

(ii) 为了使指示性词语跟其参照性词语在语义上一致，所以有时下标不一定标记在指示词之后，而是标记在整个指示性短语之后。比如，这里的指示性短语"这位与他有相同经历的农村青年"跟先行语"现在的妻子黄赛红"同指，指示性短语"这个月" 跟先行语"1996 年 7 月"同指。

(iii) 对于动词"读"来说，"电大"既像是系事，又像是方式；对于动词"考取"来说，其主体性论元既像是施事，又像是经事或感事；其客体性论元既像受事，又像是受事，又像是系事或对象。为了周全，我们把它们以析取的形式都标注上去。

(iv) 当副词"也"用以联结小句、句子和段落时，是语篇关联词语，表示并列关系(coordinate relation，简写为 COR)。

(18) {当([玉环县级行政机关$_{i}$]$_{A}${向社会}$_{L(G)}$ 公开招考[国家公务员]$_{Re}$)时}$_{T}$，[陈云峰$_{j}$]$_{P}${以([笔试]$_{Th}${全县}$_{L}$ 第二、[面试]$_{Th}${全县}$_{L}$ 第一的)成绩}$_{M}$ 被[e_i]$_{A}$ 录用，[e_j]$_{Ex}$成了[一位从事劳动仲裁工作的国家公务员]$_{Re}$。[黄赛红$_{k}$]$_{P}${也}$_{COR}${以优异的成绩}$_{M}$[被县司法局$_{l}$]$_{A}$ 录用，[e_k]$_{Ex}$[[被 e_l]]$_{A}$ 下派到[陈屿基层司法所]$_{L(G)}$，[e_k]$_{Ex}$成〈了〉$_{perf}$[一名司法员]$_{Re}$。

说明：(i) 双重方括号中 [[被 e_k]]表示承上文而省略的介词结构，为了语义理解的方便，在这里补充出来，再加上语义同指和论旨角色标记。

(ii) 当副词"也"充当语篇关联词语时，既可以用在句首，如例(17)所示；也可以用在后续句的主语之后，如本例所示。

(iii) 上面(17)(18)两个文本应该是一个段落或篇章中的，它们

通过主体性论元在两个人物之间来回变换,来交叉叙述;在表层结构上,都用了关联副词"也"来衔接叙述不同的主体的句子。在叙述时,又通过两个时间论元"1994年……,1996年"在时间上的先后关系,来显示篇章组织上采用了顺叙的方式。也就是说,这种篇章结构有比较明显的表层标志。

(19)"[气]$_{Th}$可鼓而不可泄"。[([e_i]$_{Ex}$〈曾〉$_{past}$任[省体委主任]$_{Re}$的)省委常委、常务副省长汪洋]$_A$ 提高〈了〉$_{perf}$[嗓门]$_P$ [V]:"[e_j]$_A$〈要〉$_{mod}${{以市场的眼光}$_M$ 办[体育]$_P$,[e_j]$_A$ 抓[选才]$_P$,[e_j]$_A$ 抓[教练]$_P$},[领导$_j$]$_A$〈要〉$_{mod}${[与运动员、教练员$_k$]$_D$ 滚在[一块]$_{L(G)}$,[e_{j+k}]$_A$ 艰苦奋斗,[e_{j+k}]$_A$ 进军[九运]$_{L(G)}$[e_{j+k}]$_A$ 再创[辉煌]$_R$}。"

说明:(i)"气可鼓而不可泄"和"再创辉煌"等熟语,也可以作为一个谓词性成分而不加论元角色方面的语义分析。

(ii)"的"字结构中,谓词性结构中的空语类一般跟"的"字结构所修饰的中心语同指。①

(iii)为了简单,助动词"要"等辖域中承上省去的主体性论元都可以不作补充标记,当然也可以像上面那样加上空语类标记。

(iv)标记[V]代表"说"一类被省略的言语动词。

(20)[本报]$_A${阿比让}$_L${1月5日}$_T$ 电 [记者杨贵兰]$_A$ 报道:[([e_i]$_{Ex}${在肯尼亚大选中}$_L$ 赢得[连任]$_P$ 的)肯尼亚总统莫伊$_i$]$_{Ex}${5日}$_T${在内罗毕}$_L$ 宣誓就职。{据[肯尼亚选举委员会]$_A${昨天}$_T$ 正式宣布[e_j]$_{Re}$}$_M$,{{在去年12月29日至30日举行的大选中}$_L$,[肯尼亚非洲民族联盟候选人、现任总统丹尼尔·阿拉普·莫伊$_i$]$_A${以较大优势}$_M$ 击败〈了〉$_{perf}$[14名反对党候选人]$_P$,[e_i]$_{Ex}$ 再次当选为[肯尼亚总统]$_{Re}$,[任期]$_{Th}$[V][5年]$_{Re}$}$_j$。

说明:(i)"连任"是名动词,可以作形式动词"赢得"的宾语;其

① 详见朱德熙(1978)和(1983)。

论旨角色可以归入结果,而不是受事。从意义上看,连任这种情况是人为地创造出来的。从句法形式上看,这种宾语不能作介词“把”的宾语,这跟一般的结果论元和受事论元都不同。

(ii) 隐藏在方式论元中的动词“宣布”的系事论元省去了,它跟后面的一连串小句在语义上有同指关系。

(iii)“就职、就任”等担任动词的后续句往往要交代获任的经过等情况,即整个话语采用倒叙的手法;其表层标志是使用相关的几个时间词语,比如本例中的“1 月 5 日……昨天……去年 2 月 29 日至 30 日……”等逆向性时间词系列。

(21) [罗多尔佛·塞尔特扎·塞维里诺]$_{Ex}${五日}$_{T}${在雅加达}$_{L}$ 正式就任[东南亚国家联盟(东盟)新一任秘书长]$_{Re}$。[塞维里诺$_{i}$]$_{Ex}${于去年$_{?}$ 七月}$_{T}${在吉隆坡召开的第三十届东盟外长会议$_{j}$ 上}$_{L}$ 被[e$_{j}$]$_{A}$ 任命为[东盟秘书长]$_{Re}$。[他$_{i}$ 的任期]$_{Th}$〈将〉$_{fut}${于二OO二年}$_{T}$ 结束。[塞维里诺$_{i}$]$_{Th}$是[菲律宾的一位外交家]$_{Re}$,[e$_{i}$]$_{Ex}$〈曾〉$_{past}$先后担任[菲律宾驻美国、中国和马来西亚等国的外交使节]$_{Re}$。{一九九二至一九九七年间}$_{T}$,[塞维里诺$_{i}$]$_{Ex}$任[菲外交部副部长]$_{Re}$,[e$_{i}$]$_{Ex}$负责[与东盟有关的事务等]$_{Re}$。

说明:(i) 这里第 2 句中的时间指示词“去年”的参照词要到更前面的上文(包括新闻引语)中去找。这个参照时间一般是新闻发出的时间。我们用问号表示在本段中其先行语不明确。

(ii)“就职、就任”等担任动词的后续句往往要交代获任的经过等情况,即整个话语采用倒叙的手法;其表层标志是使用相关的几个时间词语,比如本例中的“五日……去年七月……”等逆向性时间词系列。有时还要倒叙主体性论元的过去的任职经历,其表层标志是使用过去时标记“曾(经)”,然后是一组相关的顺序性的时间词语,比如本例中的“……曾……一九九二至一九九七年间……”。

(22) [尼日尔警方]$_{A}${日前$_{?}$}$_{T}$ 破获[一个([e$_{i}$]$_{A}$ 企图暗杀[迈纳萨拉总统]$_{P}$ 的)阴谋团伙$_{i}$]$_{P}$。[尼日尔通讯社]$_{A}${二日}$_{T}$报道说,[[被捕的三名团伙成员]$_{A}$ 供认,[[他们$_{i}$]$_{A}$〈原〉$_{past}$ 定

[{于去年$_?$十二月二十九日}$_T$行动]$_{Re}$，[暗杀对象]$_{Th}${除总统以外}$_{Re}$，还包括[几名政府重要成员]$_{Re}$]$_{Re}$]$_{Re}$。{据报道}$_M$，[该暗杀团伙$_i$的主谋]$_{Th}$是[哈马·阿马杜$_j$]$_{Re}$。[他$_j$]$_{Ex}$现任[尼(日尔)反对派“社会发展全国运动”总书记]$_{Re}$，[e_j]$_{Ex}${过去$_?$}$_T$〈曾〉$_{past}$担任〈过〉$_{pas\text{-}perf}$[政府总理]$_{Re}$。[警方]$_A${一日}$_T$〈已〉$_{past}$逮捕〈了〉$_{perf}$[哈马·阿马杜]$_P$，{但}$_{VER}$[他$_j$]$_A$〈不〉$_{neg}${承认[[e_j]$_{Th}$[跟这起暗杀活动]$_{Re}$有关]$_{Re}$}，[e_j]$_A$称[[自己$_j$]$_P$[被人]$_A$诬陷]$_{Re}$。

说明：(i) 这里的时间指示词“日前、去年、过去”的参照词要到更前面的上文(包括新闻引语)中去找。这个参照时间一般是新闻发出的时间。其中，“日前”是一个模糊性的时间指示词，泛指其参照时间的前几天。“过去”也是一个模糊性的时间指示词，泛指其参照时间“现任……”以前的时间。我们用问号作下标，表示该指示语的先行参照语在本段中不明确。

(ii) 为了语义标注的精细，“的”字结构用圆括号标示，其中的谓词性成分，也标记其论元角色等语义关系，其中跟中心语同指的空语类(主体论元或客体论元)也用下标标示。

(iii) “除总统以外”的论旨角色很难确定，这里姑且归入系事，因为总统也包括在暗杀的对象之中。

(iv) 为了简单，把“企图暗杀、报道说”等动词组合看作是一个动词性成分。

(v) 这里把动词“有关”所涉及的无生性的(inanimate)暗杀活动，归入系事论元。

(vi) 副词“不”是否定算子(negative operator，简写为 neg)，我们用花括号标志其辖域。

(vii) 这一段落的篇章结构比较复杂，基本的叙述方式是顺叙(……日前破获……供认……〔……〕……一日已逮捕……)，但是其中(方括号〔……〕所示)又有插叙(……主谋是……)，在插叙中因为涉及担任事件所以又引出倒叙(……现任……过去曾担任过……)。这种复杂的叙述安排，在表层结构上都有时间论元或时体算子作形式标志；并且，最后一句的主体论元转换时，用了语篇关联词语“但”

来表示转折关系(adversative relation,简写为 VER)。

(23) [邢云$_i$]$_{Ex}$,{1952 年}$_T$ 生,[e$_i$]$_{Ex}$[V][大学文化]$_{Re}$。[e$_i$]$_{Ex}$历任[内蒙古伊克昭盟副盟长、盟委副书记、盟长]$_{Re}$,[e$_i$]$_{Ex}${1996 年 10 月起}$_T$ 任[盟委书记、盟人大工委主任]$_{Re}$。

说明:(i) 这里的[V]代表“有、具有”一类隐含动词(implied verb)。

(iii) 从篇章结构的角度看,生平介绍类文本通常用顺叙的写作手法,顺次交代主体论元的出生、学习、工作、任职经历等事项;这种顺叙往往要用一组相关的顺序性的时间论元来显性地标志,比如本例中的“1952 年……1986 年 10 月起……”。

(24) [尚志派出所$_i$]$_{Th}$地处[哈尔滨繁华地带]$_L$,[全市最大的百货商店、最大的酒店和最大的菜市场]$_{Th}$都在[他们$_i$ 管辖区内]$_L$。{{1995 年}$_T${([杨光$_j$]$_{Ex}$上任)后不久}$_T$,[许多人$_k$]$_A$便主动找上[门]$_L$来,[有的$_{k1}$]$_A$ 拉[他$_j$]$_{P\&A}$合伙做[生意]$_P$,[有的$_{k2}$]$_A$〈想〉$_{mod}${[找[他$_i$]$_{P\&A}$做[靠山]$_P$]$_{Re}$}}$_l$。[对此$_l$]$_P$[杨光$_i$]$_A$一概拒绝。[十几年的从警生涯]$_{Cau}$,使[他$_i$]$_{P\&Se}$悟出[道理]$_R$:[身]$_{Th}$居[闹市]$_L$,[e$_i$]$_A$〈要〉$_{mod}${远离[灯红酒绿]$_{Re}$};[e$_i$]$_{Ex}$作为[一所之长]$_{Re}$,[e$_i$]$_A$ 更〈要〉$_{mod}${[{为全所民警}$_D$ 做出[表率]$_R$}。

说明:(i) 这里复数性的人称代词“他们”的先行词是机构名词短语“尚志派出所”。

(ii) 部分代词“有的”指人或事物中的一部分,其先行词是上文的“许多人”。两个“有的”的所指可以一样,也可以不一样;因此,在下标上加数字以示区别。

(iii) “做生意、做靠山、身居闹市、做出表率”等熟语,也可以作为一个谓词性成分而不加论元角色方面的语义分析。

(iv) 指示词“此”的先行参照词语是“1995 年……许多人……做靠山”一个大句,为了避免符号标注过于复杂,这里不用下划线,而是用花括号标志其界域。

(v) 这里的“想”表示愿望,也是助动词。

(vi) 担任动词“上任”是不及物动词,只能支配经事论元,表示

职务的系事论元只能是一种语义上的隐含性成分，于是职务及相关的机构、组织等必要信息项目都要从上下文中寻找。在本例中，首句的主体性论元“尚志派出所”表明了机构，末句的系事论元“一所之长”表明了职务。

(25) [该委员会$_i$]$_{Th}$是[全国法律硕士专业学位教育的专业性组织]$_{Re}$，[其$_i$ 主要任务]$_{Th}$是：[[e_i]$_A$ 指导、协调[全国法律专业学位教育活动]$_P$]$_{Re}$。[委员会主任]$_{Re}$[由司法部部长肖扬]$_{Ex}$担任，[委员]$_{Re}$是{在有关单位和专家推荐的基础上}$_M$，[由国务院学位委员会、国家教委和司法部]$_A$ 选聘。

说明：(i) 指示性短语“该委员会”的参照性先行词语要到上文去找，指示词“其”的参照词是包含指示词的短语“该委员会”。

(ii) 指示词“该委员会、其”的参照语都是“……委员会”，这使得它在话语中获得很高的话题性(high topicality)，成为整个段落展开叙述的线索。这是一种利用同一话题组织话语篇章的叙述手法。

(26) [新华社]$_A${北京}$_L${1 月 6 日}$_T$ 电 [中华人民共和国主席江泽民]$_A${根据全国人民代表大会常务委员会的决定}$_M$，任命[王学贤]$_{Ex}$为[中华人民共和国驻南非共和国特命全权大使]$_{Re}$。

(27) [新华社]$_A${达累斯萨拉姆}$_L${1 月 6 日}$_T$ 电[坦桑尼亚总统姆卡帕$_i$]$_A${5 日}$_T${([e_i]$_A${在内罗毕}$_L$ 祝贺[[莫伊]$_{Ex}$连任[肯尼亚总统]$_{Re}$]$_{Re}$)时}$_T$ 说，[[肯尼亚此次大选的成功]$_{Th}$表明[[肯尼亚和非洲国家]$_{Th}$〈不〉$_{neg}${需要[[别人]$_A$ 教[他们]$_D$[如何实行民主]$_P$]$_{Re}$}，[非洲人$_j$]$_{Se}$完全知道[[投谁的票]$_{Th}$符合[自己$_j$ 的利益]$_{Re}$]$_{Re}$]$_{Re}$]$_{Re}$。

说明：(i) 照应性代词“自己”在本例中有先行语，需要标注其同指关系。但是，疑问代词“谁”在这里是一种任指用法，没有先行语，所以不需要标注其同指关系。

(28) {1938 年 2 月}$_T$，[中共晋冀豫省委]$_A${在太岳区沁县}$_L$ 设立[办事处]$_R$，[由省委统战部部长安子文]$_{Ex}$兼任[办事

处主任]$_{Re}$。{1938 年 7 月}$_T$，{在沁县办事处的基础上}$_M$成立〈了〉$_{perf}$[中共太岳特委]$_R$，[安子文同志]$_{Ex}$任[书记]$_{Re}$，[太岳区的党组织]$_{Th}$进入〈了〉$_{perf}$[大发展阶段]$_{L(G)}$。{1939 年 4 月}$_T$，[中共太岳特委]$_{Th}$改称[中共太岳地委]$_{Re}$，[安子文同志]$_{Ex}$任[书记]$_{Re}$。{1939 年 11 月}$_T$，成立〈了〉$_{perf}$[(以[薄一波同志]$_{Ex}$任[书记]$_{Re}$的)晋东南军政委员会]$_R$。{1940 年$_j$1 月}$_T$，[中共北方局]$_A$决定[[太岳地区$_i$]$_{Th}$成为[独立的战略区]$_{Re}$，[e_i]$_A$成立[太岳区党委]$_R$，[安子文同志]$_{Ex}$任[书记]$_{Re}$]$_{Re}$。{同年$_j$1 月 19 日}$_T$，[陈赓同志]$_A$奉[八路军总部命令]$_P$率领[八路军三八六旅]$_{P\&A}$进驻[太岳区]$_L$。[晋东南军政委员会]$_{Th}$改称为[太岳军政委员会]$_{Re}$，[薄一波同志]$_{Ex}$任[书记]$_{Re}$，[主要成员]$_{Th}$有[陈赓、安子文等同志]$_{Re}$。

说明：(i)“成立太岳区党委”前省去的施事似乎是前面句子中的“太岳地区”。“以薄一波任书记”似乎不太通顺，更合适的表达是“以薄一波为书记”。

(ii) 用相关的顺序性的时间词语作一组相关句子的时间论元，是顺叙手法展开叙述、组织篇章的典型手段。比如本例中的“1938 年 2 月……1938 年 7 月……1939 年 4 月……1939 年 11 月……1940 年 1 月……同年 1 月 19 日……”。

(29) [何长工]$_{Ex}${1952 年 8 月$_i$}$_T$调入[地质部]$_L$。{此前$_i$}$_T$[e_i]$_{Ex}$〈曾〉$_{past}$任[重工业部副部长、代部长]$_{Re}$，{以航空、钢铁、造船、电机和动力工业为重点}$_M$，[e_i]$_A$抓〈了〉$_{perf}$[重工业部的组建工作]$_P$，[e_i]$_A$奠定〈了〉$_{perf}$[我国重工业和航空工业发展的基础]$_P$。

说明：(i) 指示词“此前”的参照词语是“1952 年 8 月”，所指为：1952 年 8 月以前的一段时间。这说明指示词的所指跟参照词语的所指有关，但不一定相同。

(ii)“就职、就任、调任、调入”等担任动词的后续句往往要交代获任的经过等情况，即整个话语采用倒叙的手法；其表层标志是使用相关的几个时间词语，有时还要倒叙主体性论元过去的任职经历，其

表层标志是使用过去时标记“曾(经)”等时体算子,然后是一组相关的顺序的时间词语。

(30) [艾哈迈德·本·穆罕默德·萨利姆]$_{A}$提交的(?了)[理事会年度报告]$_{P}$,[与会代表$_{i}$]$_{A}$[对萨利姆在执行秘书处1997年工作计划中所做的努力和取得的成绩]$_{Re}$表示[感谢]$_{Re}$,[e_i]$_{A}${并}$_{ADD}$一致选举[萨利姆]$_{Ex}$继续担任[下一届理事会秘书长]$_{Re}$。

说明:(i) 这里“提交的”中的“的”疑为“了”之误。

(ii) 这里的空语类[e_i]如果用其先行语“与会代表”代进去以后,句子就不通了;但是,从语义的论元结构和句法的深层结构的角度看,这里的确有一个空缺的成分,其论旨角色是施事。这就是篇章中的句子的结构特点:通过省略来衔接句子。

(iii) 这里把“继续担任”当作一个动词性成分而不加语义关系分析,也可以把“继续”处理为状语,跟“一致”一样不作语义关系分析。

(iv) 语篇关联词语“并”把最后两个小句衔接起来了。

(31a) [新华社]$_{A}${斯德哥尔摩}$_{L}${1月7日$_{i}$}$_{T}$电(记者许福瑞[V])[([e_j]$_{A}$〈曾〉$_{past}${为调解巴勒斯坦和以色列冲突}$_{Ai}$作出〈过〉$_{past\text{-}perf}$[努力]$_{R}$ 的)挪威政府$_{j}$]$_{A}$,{最近$_{i}$}$_{T}$任命[罗德—拉森$_{k}$]$_{Ex}$为[驻中东巡回大使]$_{Re}$,[e_k]$_{A}$重新负起[([e_k]$_{A}$调解[巴以冲突]$_{P}$ 的)使命]$_{P}$。

说明:(i) 隐藏在主体论元的修饰语中的过去时标记“曾”,使得整个句子在叙述方式具有一种隐性的顺叙色彩。如果改成显性的顺叙表达,那么将可能是下面的(31b)。

(31b) [挪威政府$_{i}$]$_{A}${过去$_{j}$}$_{T\&TEM\text{-}i}$〈曾〉$_{past}${为调解巴勒斯坦和以色列冲突}$_{Go}$作出〈过〉$_{past\text{-}perf}$[努力]$_{R}$,[e_i]$_{A}${最近$_{j}$}$_{T\&TEM\text{-}i}$任命[罗德—拉森$_{k}$]$_{Ex}$为[驻中东巡回大使]$_{Re}$,[e_k]$_{A}$重新负起[([e_k]$_{A}$调解[巴以冲突]$_{P}$ 的)使命]$_{P}$。

说明:(i) 顺序的时间论元和过去时标志“过去曾……最

近……”，使得整个句子在叙述方式具有一种显性的顺叙色彩。

(32) [今年51岁、有34年军龄的齐文明$_{i}$]$_{A}${也}$_{COR}${为此$_{?}$}$_{Rn}$赢得〈了〉$_{perf}$[很多荣誉]$_{R}$，[e_{i}]$_{A}$荣立[二等功、三等功]$_{R}$，[这项成果]$_{P}$被[$e_{?}$]$_{A}$评为[国家科技进步三等奖]$_{Re}$。[他$_{i}$]$_{P}$被[$e_{?}$]$_{A}$评为[空军科技先进个人]$_{Re}$，[e_{i}]$_{P}${并}$_{ADD}$被[$e_{?}$]$_{A}$评为[军内有突出贡献的专家]$_{Re}$，[e_{i}]$_{Ex}$成为[北方心脑血管病血流变学会副主任委员]$_{Re}$。

说明：(i) “被”引导的施事论元省略了，并且在本段中没有出现其先行语，所以用问号标注。必须说明的是，这些空语类都用问号作下标，但并不表示它们是同指的；也就是说，它们完全可以跟不同的先行语发生语义同指关系。

(ii) 第二句中的代词“他”跟第一句中的先行词“齐文明”之间的照应关系，正好起到语篇连贯作用。

(33) [南非真相委员会副主席伯瑞恩$_{i}$]$_{A}${同日$_{?}$}$_{T}$呼吁，[[他$_{i}$]$_{Se}$希望{[[博塔$_{j}$]$_{A}$〈能〉$_{mod}${{在最后一刻}$_{T}$改变[态度]$_{P}$，[e_{j}]$_{Se}$同意[[e_{j}]$_{A}$到[真相委员会]$_{L}$作证]$_{Re}$}]$_{Re}$}$_{k}$]$_{Re}$。{果真如此$_{k}$}$_{SUP}$，[真相委员会]$_{A}$〈将〉$_{fut}$建议[[卡恩]$_{A}$撤诉]$_{Re}$。[博塔$_{j}$]$_{Ex}$〈曾〉$_{past}$任[旧南非的国防部长、总理和总统]$_{Re}$，[他$_{j}$掌握的重要情况]$_{Th}${对真相委员会完成其使命}$_{Re}$至关重要。[真相委员会]$_{A}$〈曾〉$_{past}$ 3次传唤[[他$_{j}$]$_{A}$到场听证]$_{Re}$，{但}$_{VER}$[他$_{j}$]$_{A}$均{以有病等理由}$_{Rn}$拒绝[出席]$_{P}$，[e_{j}]$_{A}${并}$_{ADD}$指责[[真相委员会的工作]$_{Th}$是[“马戏团表演”和“政治迫害”]$_{Re}$]$_{Re}$。

说明：(i) 为了简单，像“果真如此、但、并”等承接性语篇关联词语，直接在其后标注。其中，“果真如此”表示假设关系(suppositional relation，简写为SUP)。碰到“因为……所以……、如果……那么……、只要……就……”等配套的表示条件的关联词语，则在其所关联的小句或句子之后标注语义关系，关联词语下面加着重点。其中，指示词“此”参照词语即是动词“希望”的系事“博塔……作证”。为了清晰，不用下划线，而是用花括号标记范围。

(ii) 这一段落通过主体性论元在两个人物及其所属的组织之间

来回变换，来交叉叙述。在叙述时，又通过过去时标志“曾”，来显示篇章组织上的插叙，即补充交代为什么需要博塔出来作证。同时，用了“果真如此、但、并”等承接性语篇关联词语来衔接句子。

(34) [摩洛哥新一届两院制议会]$_A${7 日$_i$}$_T$选出[第一任参议院议长]$_{Re}$，[原摩洛哥一院制议会议长艾赛义德]$_{Ex}$当选为[第一任议长]$_{Re}$。[代表院议长$_j$]$_{Re}$〈已〉$_{past}${于 6 日}$_T$选举产生，[原议会第一副议长拉迪]$_{Ex}$当选[e_j]$_{Re}$。{至此$_i$}$_T$，[摩洛哥议会从一院制向两院制的转变]$_{Th}$宣告[完成]$_{Re}$。[现年 60 岁的艾赛义德$_k$]$_{Ex}$是[摩(洛哥)右翼党派宪政联盟的议员]$_{Re}$，[e_k]$_{Ex}$[V][法学博士]$_{Re}$，[e_k]〈曾〉$_{past}$担任〈过〉$_{past\text{-}perf}$[国务秘书和阿拉伯议会联盟委员会主席等职]$_{Re}$。

说明：(i) 在时间论元“至此”中，指示词“此”有两个候选的参照词语“7 日”和“6 日”；但是，根据文意来推算，应该选择比较靠后的时间，即“7 日”。

(ii) 本例通过“7 日……6 日……”等逆向性时间词系列，显示整个话语采用了倒叙的手法。同时，担任动词句往往还要交代主体论元现在所担任的其他职务，有时还要倒叙主体性论元过去的任职经历，其表层标志是使用过去时标记“曾(经)”。

(35) [墨西哥总统塞迪略]$_A${7 日$_i$}$_T$任命[女参议员罗萨里奥·格林]$_{Ex}$为[外交部长]$_{Re}$。[格林]$_A${同日$_i$}$_T${在就职后}$_L$宣布，[[墨西哥外交部]$_A$〈将〉$_{fut}${本着尊重主权，不干涉别国内政的外交政策}$_M$，发展[与各国的关系]$_{Re}$]$_{Re}$。[格林]$_A${在谈到与亚太国家关系时}$_T$指出，[[亚太国家]$_{Th}$具有[巨大的投资、贸易与合作潜力]$_{Re}$，[墨西哥]$_A$〈将〉$_{fut}$加强[与该地区的合作]$_{Re}$]$_{Re}$。[格林$_j$]$_{Th}${1941 年}$_T$生于[墨西哥城]$_L$。[她$_j$]$_{Th}$是[墨西哥历史上第一位女外交部长]$_{Re}$。

说明：(i) 本例通过“7 日任命……同日…就职……”这种时间指示词跟先行语之间的参照关系、任命事件跟担任事件之间的先后事件关系，来组织话语篇章。同时，担任动词句往往还要交代主体论元的生平，其表层标志是使用时间词语作时间论元或者使用时体算

子作标记，本例用的是显性的时间词语“1941年”。

(36) [冈崎嘉平太$_i$，这位$_i$([e_i]〈曾〉$_{past}$任[日本全日空航空公司总裁]$_{Re}$的)老人]$_A$，{生前}$_T$一百多次来[中国]$_{L(G)}${为恢复中日邦交}$_{Ai}$奔波。{与周恩来$_k$交往没多久$_j$}$_T$，[这位中国总理]$_{Th}$就成〈了〉$_{perf}$[他$_i$心中所崇敬的孔子般的“圣人”]$_{Re}$。{从此$_j$}$_T$[他$_i$]$_A$[将这位“圣人”$_k$的相片]$_P$虔敬地珍藏在{胸前贴身处}$_{L(G)}$，{几十年}$_T$一如既往{直至离开人世}$_T$依然不舍。{当([他$_i$的儿子$_l$]$_A${向摄制组}$_D$讲起[老人去世的情景]$_{Re}$)时}$_T$，[e_l]$_{Se}$十分平静(，[e_l]$_A$[V])：[“[父亲$_i$]追随[他$_i$崇敬的人]$_{Re}$到[天国]$_L$去〈了〉$_{perf}$，[e_i]$_{Se}$是幸福的。”

说明：(i) 为了简单，把复指性成分处理成一个论元。“一百多次、虔敬地”等状语性修饰成分，不作语义标注。

(ii) 为了语义显豁，把说明叙述文本中经常省略的“说”一类言语动词，用[V]的形式标记出来。

(iii) 像“父亲”等有价名词，是一种变量性成分，其语义所指要依赖其从属论元作为参照成分来确定。即要让人知道这“父亲”指的是谁的父亲。我们约定：当这种变量性成分的上下文有同指的成分出现时，以相同的下标来直接标记其所指；当这种变量性成分的上下文没有同指的成分出现时，以相同的下标加百分号%来间接标记它跟其参照成分的语义从属关系。

(iv) 为了简单，语气词“了”也标记为表示完成体的虚词。

(v) 后续句中的代词“他”跟第一句中作主体性论元的先行词“冈琦嘉平太”之间的照应关系，正好起到语篇连贯作用。并且，正是这种凝聚力很强的照应和衔接关系，使得这个先行语成为整个段落的话题。

(37) a. [童志成$_i$]$_{Th}$来不及[[e_i]$_A$回给[陈隆年]$_D$[一句俏皮话]$_P$]$_{Re}$，[公司其他领导梁德妍、王润培、叶石池、李广尊等$_j$]$_{Se}$听说[[他$_i$]$_A$出差回来〈了〉$_{perf}$]$_{Re}$，[e_j]$_A${也}$_{COR}$都不约而同地来见见[([上级]$_A$〈刚〉$_{past}$任命[e_i]$_{P\&Ex}$当[一把手]$_{Re}$的)童志成]$_{Ta}$。

说明：(i) 为了简单，这里把时间副词"刚"也标记为表示过去时的虚词。

(ii) 这一段落通过主体性论元在一个人跟一组人之间变换，来交叉叙述。

(iii) 上面把"来不及"当作一个主要动词来分析，如果把它看作是助动词，那么可以作如下分析：

(37) b. [童志成$_i$]$_A$〈来不及〉$_{mod}${回给[陈隆年]$_D$[一句俏皮话]$_P$}，[公司其他领导梁德妍、王润培、叶石池、李广尊等$_j$]$_{Se}$听说[[他$_i$]$_A$ 出差回来〈了〉$_{perf}$]$_{Re}$，[e_j]$_A${也}$_{COR}$都不约而同地来见见[([上级]$_A$〈刚〉$_{past}$任命[e_i]$_{P\&Ex}$当[一把手]$_{Re}$的)童志成]$_{Ta}$。

说明：(i) 相比之下，把"来不及"处理为助动词的分析更简单，并且正确地反映出整个句子所表示的时间的非现实性。

(38) {1995 年 6 月}$_T$，[杨丽华$_i$]$_{Ex}$[被中国国际航空公司]$_A$任命为[飞行总队副总队长]$_{Re}$。[她$_i$]$_A${从一个普通的空中小姐}$_{L(S)}$，{在空中工作 22 年后}$_T$ 走上〈了〉$_{perf}$[领导岗位]$_{L(G)}$，[e_i]$_{Th}$不容易啊！

说明：(i) 这里"不容易"的省略的主语既可以是"杨丽华"，也可以是整个句子"她……走上了领导岗位"。关于这种代词和空语类在所指上的波动现象，请看袁毓林(2002c)。

(ii) "就职、就任、调任、调入、选举、任命"等担任动词和任命动词的后续句往往要交代获任的经过或生平事迹等情况，即整个话语采用倒叙的手法。在本例中，第二句中作主体性论元的代词"她"跟第一句中作主体性论元的先行词"杨丽华"之间的照应关系，正好起到语篇关联作用。

(39) [新华社]$_A${大连}$_L${1 月 10 日}$_T$ 电[大连市第十二届人民代表大会第一次会议$_i$]$_A${1 月 10 日}$_T$ 选举[于学祥]$_{Ex}$为[市人大常委会主任]$_{Re}$，[e_i]$_A$ 选举[薄熙来]$_{Ex}$为[大连市市长]$_{Re}$。

(40) [墨西哥恰帕斯州议会$_i$]$_A${7 日}$_T$ 批准[[胡里奥·鲁

依斯]$_{Ex}$辞去[州长职务]$_{Re}$]$_{Re}$,{并}$_{ADD}$[e_i]$_A$ 任命[众议员罗伯特·阿尔沃雷斯·纪廉]$_{Ex}$为[新州长]$_{Re}$。{自({去年12月22日}$_T${在恰帕斯州$_j$ 切纳洛市}$_L$ 发生[([45名印第安人]$_P$ 被[$e_?$]$_A$ 杀)惨案]$_{Re}$)后}$_T$,[([该州$_j$ 和全国各地]$_A$ 要求[[鲁依斯]$_{Ex}$引咎辞职]$_{Re}$的)呼声]$_{Th}$愈来愈高。[鲁依斯的辞职$_k$]$_{Th}$受到[各方面欢迎]$_{Re}$。[一些人士]$_{Se}$认为,[[这样做$_k$]$_{Th}$〈可以〉$_{mod}${[对公正解决屠杀惨案,推动政府与该州游击队组织萨帕塔民族解放军恢复和谈]$_{Re}$,产生[积极影响]$_R$}]$_{Re}$。

说明:(i)“发生”等表示存现、消失的动词之后的论元,其论旨角色暂时归入系事,以便跟动词前可以出现主体性论元的句子相协调。例如:“他们发生了一点儿误会 ~ 误会最终发生了”。

(ii) 为了语义所指上的一致,在整个包含指示词的短语“这样做”之后加同指下标。

(iii)“就职、调任、选举、任命、免去”等担任动词和任免动词的后续句往往要交代获任的经过或生平事迹等情况,或者是被免的原因或被免者的生平事迹等情况,即整个话语采用倒叙的手法。

(41){深化改革的大潮中}$_T$,[知识分子韩兵$_i$]$_A$ 毅然辞去[公职]$_{Re}$回到[生他$_i$ 养他$_i$ 的鄱阳湖畔]$_{L(G)}$干起〈了〉$_{perf}$[农业立体开发]$_P$。{经过几年奋斗}$_M$,[他$_i$]$_A$ 带领[乡亲们$_j$]$_P${通过股份合作制}$_M$,[e_{i+j}]$_A$ 办起[集农、科、贸一体的农业集团公司]$_R$,[e_{i+j}]$_A$ 实现〈了〉$_{perf}$[农业产业化]$_R$。{艰苦的创业中}$_T$,[韩兵$_i$]$_A${在乡亲的支持下}$_M$ 克服[种种困难]$_P$,[e_i]$_A${并}$_{ADD}$赢得〈了〉$_{perf}$[女主人公的爱情]$_P$。{该剧$_?$}$_{TOP}${[情节]$_{Th}$曲折,[感情纠葛]$_{Cau}$让[人]$_{P\&Se}$回肠荡气}$_{COM}$。

说明:(i) 指示性短语“该剧”的参照性词语要到更前面的上下文中去找。

(ii) 从论旨角色关系上看,话题性成分“该剧”跟其说明部分中的谓词性成分“曲折”和“让……回肠荡气”都没有关系。碰到这种情况,索性标记话题(topic,简写为 TOP)和说明(comment,简写为 COM)这种语用和话语结构关系。

(iii)“该剧”是一价名词“情节”的从属成分，为了简单，就不标记这种降级述谓关系了。

(42)［诸葛彩华$_{i}$］$_{Ex}$［从县委副书记岗位］$_{L(S)/Re1}$调任［代县长］$_{L(G)/Re2}$。［上苍］$_{A}$似乎故意地考验［她$_{i}$］$_{P}$。［海岛］$_{Th}$缺［水］$_{Re}$，{偏偏}$_{VER}${1995 下半年}$_{T}${玉环岛上}$_{L}$［滴水］$_{Th}$〈未〉$_{neg}${下}，［旱灾］$_{Th}$严重，［生活用水］$_{Th}$频频告急。

说明：(i) 对于动词“调任”来说，原来的职务和后来的职务都是系事；同时，从路径隐喻(path metaphor)的角度来看，原来的职务是源点，后来的职务是终点。为了周全，把这两套语义角色的名目都标注上去，并用析取号/来标记。

(ii) 语气副词“偏偏”在这里表示转折关系，有语篇衔接功能。所以需要标注。

(43){经查}$_{M}$，［张斌昌$_{i}$］$_{A}${在(［e$_{i}$］$_{Ex}$任［酒泉钢铁集团公司副总经理］$_{Re}$)期间}$_{T}$，利用［职权］$_{P}$，［e$_{i}$］$_{A}$多次收受［他人贿赂的现金、国库券等折合人民币 80 多万元］$_{P}$。{1995 年$_{j}$ 8 月}$_{T}$，{(［张斌昌$_{i}$］$_{Ex}$调任［兰州钢铁集团总经理］$_{Re}$)后}$_{T}$，［e$_{i}$］$_{A}${于同年$_{j}$ 11 月}$_{T}$［向某外方合资企业董事长］$_{D}$索要［［人民币］$_{Th}$［V］［5 万元］$_{Re}$］$_{P}$。

说明：(i) 为了简单，我们假定“人民币 5 万元”中隐含了“达”一类动词，并据此标定“人民币”和“5 万元”的论旨角色。

(44)［史有彪］$_{Ex}${1987 年 8 月}$_{T}$调任［市委农工部副部长、农业委员会副主任］$_{Re}$。{在多年的工作中}$_{T}$，［史有彪$_{i}$给人的印象］$_{Th}$是［兢兢业业］$_{Re}$，［e$_{i}$］$_{Th}$是［个出了名的“老实部长”］$_{Re}$。

说明：(i) 这里的成语“兢兢业业”是谓词性的，在论旨角色上可以归入系事。

(45a)［邱娥国$_{i}$的职务］$_{Ex/Th}${虽}$_{CES\text{-}i}$〈已〉$_{past}$升为［分管户籍、外勤的副所长］$_{Re}$，{但}$_{VER\text{-}i}$［他$_{i}$］$_{A}$还是{“按照原来的那样}$_{M}$做［PRO］$_{P/R}$”，［e$_{i}$］$_{A}$经常深入［辖区］$_{L}$，［e$_{i}$］$_{A}${为实现辖区发案

少、秩序好、群众满意}$_{Ai}$而努力。{1997 年}$_{T}$，[邱娥国]$_{A}$收到[4000 多封群众来信]$_{P}$。[对于自己$_{i}$职责范围之内又能办得到的事]$_{Re}$，[邱娥国$_{i}$]$_{A}$都尽力去做。

说明：(i) 动词“做”后面省去了受事或结果一类客体论元。

(ii) 这里的时间副词“已”确实是一个表示过去时的虚词。

(iii) 表示转折关系的连词“虽……但……”是篇章关联词语。对于这种配套的有条件关系的连词，一种标记法是直接把它们标记为篇章关联词语，并加上相同的下标，以示它们之间的条件关系；另一种标记法是分别在这种关联词语下面加着重点，把它们关联的小句或句子用花括号套起来，并在其后标注条件(CON)和结果(CSQ)等逻辑语义关系，如下面所示：

(45b){[邱娥国$_{i}$的职务]$_{Ex/Th}$虽〈已〉$_{past}$升为[分管户籍、外勤的副所长]$_{Re}$，}$_{CES\text{-}i}${但 [他$_{i}$]$_{A}$ 还是{“按照原来的那样}$_{M}$做[PRO]$_{P/R}$”，[e$_{i}$]$_{A}$经常深入[辖区]$_{L}$，[e$_{i}$]$_{A}${为实现辖区发案少、秩序好、群众满意}$_{Ai}$而努力。}$_{VER\text{-}i}${1997 年}$_{T}$，[邱娥国]$_{A}$收到[4000 多封群众来信]$_{P}$。[对于自己$_{i}$职责范围之内又能办得到的事]$_{Re}$，[邱娥国$_{i}$]$_{A}$都尽力去做。

(46a)[利维$_{i}$]$_{A}${当天$_{?}$下午}$_{T}${在特拉维夫}$_{L}$宣布，[{由于[内塔尼亚胡]$_{A}$〈未〉$_{neg}${〈能〉$_{mod}${[对他$_{i}$所提出的一些修改 1998 年度国家预算的要求]$_{Re}$作出[答复]$_{R}$}}}$_{Rn}$，[他$_{i}$]$_{A}$决定[[e$_{i}$]$_{Ex}$辞去[外长职务]$_{Re}$]$_{Re}$]$_{Re}$。

说明：(i) 指示词“当天”参照词语要到上文中去找。

(ii) 因为“由于”引导的原因小句及相应的结果小句联合起来作动词“宣布”的系事宾语，并且结果小句又没有出现连词；所以，可以简单地把原因小句处理为原因论元。当然，也可以分别处理为原因小句和结果小句，作如下这种标注处理：

(46b)[利维$_{i}$]$_{A}${当天$_{?}$下午}$_{T}${在特拉维夫}$_{L}$宣布，[{由于[内塔尼亚胡]$_{A}$〈未〉$_{neg}${〈能〉$_{Mod}${[对他$_{i}$所提出的一些修改 1998 年度国家预算的要求]$_{Re}$作出[答复]$_{R}$}}}$_{CAS}$，{[他$_{i}$]$_{A}$决定

[[e_i]$_{Ex}$辞去[外长职务]$_{Re}$]$_{Re}$}$_{CSQ}$]$_{Re}$。

说明：(i) 在这里，原因小句(cause clause)记作 CAS，结果小句(consequence clause)记作 CSQ。

(ii) 这里的第一个人称代词"他"的先行语不是前面邻近的人名"内塔尼亚胡"，而是句首的人名"利维"。这种代词照应的求解规则，要用到高话题性这一概念。可以大概地表述如下：如果代词有几个候选的先行语，那么高话题性的成分优先。至于话题性的高下，至少有语法地位、语法位置、有定性等指标。一般地说，主语＞宾语，主句＞从句，邻近＞远程，有定＞无定，旧信息＞新信息，等等。在本例中，"利维"是主句主语，"内塔尼亚胡"是从句主语；但是，后者更邻近"他"。因此，事实上，如果纯粹从结构上看，"他"在照应关系上是有歧义的。比如，只要把"答复"改成"说明"，再把主句的宾语从句作一些调整，就可以让"他"回指"内塔尼亚胡"。例如：

(46c)[利维$_i$]$_A${当天，下午}$_T${在特拉维夫}$_L$ 宣布，[{由于[内塔尼亚胡$_j$]$_A$〈未〉$_{neg}${〈能〉$_{Mod}${[对他$_j$ 所提出的一些修改 1998 年度国家预算的要求]$_{Re}$作出[说明]$_R$}}}$_{CAS}$，{[他$_i$]$_A$ 决定[[e_i]$_{Ex}$搁置[这项要求]$_P$]$_{Re}$}$_{CSQ}$]$_{Re}$。

说明：(i) 在这里，两个"他"分别跟不同的人名发生回指关系：从句中的"他"跟从句中的人名发生回指关系，主句中的"他"跟主句中的人名发生回指关系。

(47) {1992 年}$_T$，[刘涛$_i$]$_A$ 进入[江西农用机械厂]$_L$，[e_i]$_{Ex}$成为[总工程师]$_{Re}$。

2　新闻短讯的语义标注

这些文本都是新闻报道中的比较完整的短讯，或者是完整的报道中摘取出来的句子，主题都是关于职务变动的。来自网上检索到的关于职务变动的新闻报道。这些文本的编号原文没有，是我们为了查找和核对的方便而加上去的。

(1) [原中联办主任姜恩柱]$_{Ex}$获任[人大外事委副主任委员]$_{Re}$。(南方网)

(2) [现任信息产业部部长王旭东]$_{Ex}$接任[国信办主任一职]$_{Re}$。(南方网)

(3) [黄卫$_i$]$_{Ex}$任[建设部副部长]$_{Re}$,[e_i]$_{Ex}$〈不〉$_{neg}$再担任[江苏省副省长职务]$_{Re}$。(南方网)

(4) {9月7日}$_T$,[铁道通信信息有限责任公司]$_{Th}$发生[重大人事变动]$_{Re}$。[原总经理彭朋]$_{Ex}${经[董事会$_i$]$_A$召开[临时会议]$_R$}$_M$,被[e_i]$_A$解除[职务]$_{Re/P}$,[新任铁通公司总经理]$_{Re}$[由乔金洲]$_{Ex}$担任。(新浪网)

说明:(i) 为了简单和方便,可以把"解除职务"看作是一个动词性成分而不加分析。因为,在结构和语义上,述宾结构"解除职务"大致相当于动词"解职"。当然,为了分析的彻底,也可以把"职务"看作是一个比较抽象的系事论元。这样,可以使论元结构关系的分析真正地落实到以动词为单位,并且是以动词为中心。

(ii) 对于动词"解除"来说,"职务"的论旨角色是受事;但是,为了跟担任动词的客体论元(其论旨角色为系事)协调,也可以标定为系事。"解除"是个双宾语动词,在本例中与事宾语通过被动化而前置于动词,受事(或系事)宾语留在原位。

(5) [驻港部队司令员$_i$]$_{Th/P/Re}$调整,[熊自仁]$_{Ex}$离任,[王继堂]$_{Ex}$接任[e_i]$_{Re}$。(南方网)

说明:(i)"驻港部队司令员"对于"调整"来说,其论旨角色可以笼统地归入主事。但是,如果追求语义角色关系的精密;那么"调整"应该有一个施事,在本例中被省略了,参看下面例(12)。于是"调整"的对象"驻港部队司令员"是客体论元,可以归入受事。当然,为了在整个句子中使其语义角色相对统一,也可以归入系事。

(ii)"离任"的意思是离开原来担任的职务,即不再担任这一职务;是一个不及物动词,在结构上不能带宾语。但是,它在语义上确实隐含了一个系事性成分,在这里是"驻港部队司令员"。而"接任"的意思是接替职务,是一个及物动词;在结构上省略了一个系事宾

语，在这里是“驻港部队司令员”。可见，意义相同或相反的动词，其配位能力有时差别很大。

(6)［中共］$_{A}$ 免去［张文康］$_{Ex}$［卫生部党组书记职务］$_{Re}$，［高强］$_{Ex}$继任。（南方网）

说明：(i)“继任”的意思是接替前任的职务，跟上例中的“离任”一样，是一个不及物动词，在结构上不能带宾语。但是，在语义上确实隐含了一个系事性成分，在这里是“卫生部党组书记职务”。

(7)［原北大副校长陈章良］$_{Ex}$任［中国农大校长］$_{Re}$。（南方网）

说明：(i) 这里通过经事论元中的修饰语，表示这不是一个简单的担任事件，而是一个由离任和新任两个事件复合成的调任事件。

(8)［人大常委会］$_{A}$ 通过［一批免职与任命名单］$_{P}$。（南方网）

说明：(i) 在本例中，动词“免职”和“任命”都是名动词，直接修饰名词，表示一种属性；因此，不必为它们标注论旨角色关系。

(9)［曾庆红$_{i}$］$_{A}$ 接替［胡锦涛］$_{P}${［e_{i}］$_{Ex}$ 兼任［中央党校校长］$_{Re}$}。（南方网）

说明：(i) 这种标注方式，是认为连谓结构的后段省略或隐含了一个跟前段的主语同指的经事成分。当然，为了简单，对这种中间不用逗号点断的连谓结构，当其中不同的动词性成分的主体论元相同时，后续动词性成分空缺的主体论元可以不作标记。即完全可以由某种缺省性句法、语义规则来处理。

(10)［新当选的中央政治局常委］$_{A}$［与中外记者］$_{D}$ 见面。（南方网）

说明：(i) 准二元动词通常用“和、同、跟、与”引进其与事论元。

(11)［陕西体(育)彩(票)中心领导班子］$_{P}$ 被［$e_{?}$］$_{A}$ 勒令辞职。（南方网）

说明：(i) 为了简单和方便，“勒令”跟其后面的动词可以看作是一个动词性的结构而不加分析。

(ii) 我们用[$e_?$]标记在本段中先行语不明确的空语类。

(12) [国务院]$_A$ 调整[三峡工程建设委员会]$_P$，[温家宝]$_{Ex}$ 兼任[主任]$_{Re}$。(南方网)

3　新闻全文的语义标注

这些文本都是新闻报道的全文，来自网上检索到的关于职务变动的新闻报道。这些文本的编号和标题前的“(副)标题(1/2)”字样，原文没有，是我们为了查找和核对的方便而加上去的。

(1) 标题：[俄罗斯第一副外长]$_{Th/P/Re}$ 易人[阿夫杰耶夫]$_P$ 被[e_j]$_A$ 解除[职务]$_{Re}$

[新华网]$_A${莫斯科}$_L${2 月 23 日$_i$}$_T$ 专电[俄罗斯总统普京$_j$]$_A${23 日}$_T$ 签署[命令]$_R$，[e_j]$_A$ 解除[亚历山大·阿夫杰耶夫的第一副外长职务]$_{P/Re}$，{同时$_i$}$_T$ 任命[瓦列里·洛希宁]$_{Ex}$ 为[第一副外长]$_{Re}$。

[俄塔(斯)社 k]$_A$ 援引[总统新闻秘书格罗莫夫的话]$_P$ [e_k]$_A$ 说，[阿夫杰耶夫]$_{Th}$ 另有任用。{据可靠人士透露}$_M$，[阿夫杰耶夫]$_{Ex}$〈将〉$_{fut}$ 出任[俄罗斯驻法国大使]$_{Re}$。(新华网)

说明：(i) 为了简单，我们把“易人”看作是一个动词，其意义是：(职务等)更换担当者；类似于动词“易手”，其意义是：(政权、财产等)更换占有者。另外，“易人”的意思跟第 2 部分例(5)(12)中的“调整”相似，但是“调整”是及物动词，各从属成分的论旨角色关系比较明显；而“易人”是不及物动词，其主语的论旨角色很不明朗，既像是主事、又像是受事或系事。这里为了周全，以析取形式都标注上去了。

(ii) 根据《现代汉语词典》，“专电”的意思是：记者专为本报社报道新闻而由外地用电话、电报、电传发来的稿子，区别于通讯社供稿。因此，“新华网莫斯科 2 月 23 日专电”的意思是：新华网记者从

莫斯科于2月23日发出的专电。为了标注的简单和语义角色关系的清晰，我们暂时忽略其中的名词化之后的指称性意义，而径直标注其名词化之前的谓词性结构的陈述性意义关系。也就是说，姑且把这里的“专电”解释为“发出专电”。参见第1部分第(1)例的说明中对“电”的解释。

(iii)“亚历山大·阿夫杰耶夫的第一副外长职务”对于动词“解除”来说，其论旨角色应该是受事；但是，为了跟其他职务变更动词的论元的论旨角色相对应和一致，这里同时标注上系事这种角色。

(iv)“同时”是一个时间指示词，其更完整的形式是“与此同时”；因此，应该标注其参照关系。参看第7例第7段开头的“与此同时”。

(v)我们把“另有安排”一类动词性结构看作一个整体而不加语义关系分析。

(vi)从篇章结构上看，这篇报道分为上下两个段落，第一段有三个小句：第一小句总提“签署命令”，接下来两个小句说明命令的具体内容是任免(……解除……，……任命……)；其中，最后一个小句用时间指示词“同时”来充当时间论元，并起到衔接上下文的作用。第二段交代被免职者的出路，分别用两个句子，从不同的消息渠道由模糊到清晰地说明被免职者“另有任用”和任用的机构及其职务(驻法大使)。

(2) 标题：{“宝马”假彩案}$_{TOP}$：{[陕西体彩中心主任贾安庆]$_{P}$ 被[e_j]$_{A}$ 撤职}$_{COM}$

[中新网]$_{A}${5月11日$_{i}$}$_{T}$ 电 {据央视国际消息}$_{M}$，[陕西省体育局$_{j}$]$_{A}${今天$_{i}$}$_{T}$[对陕西体(育)彩(票)中心主任贾安庆]$_{D/Ex}$ 作出[撤职的决定]$_{R}$。

[([贾安庆]$_{P/Ex}$被[e_j]$_{A}$ 撤职的)原因]$_{Th}$是[{3月23日}$_{T}${在([西安市即开型体育彩票]$_{P}$ 发售)过程中}$_{T}$，{由于[体彩部门$_{k}$]$_{A}$ 用[人]$_{P}$]$_{Th}$失察、[[e_k]$_{A}$ 监管[PRO]$_{P}$]$_{Th}$不力、{加之}$_{ADD}$[相关法律]$_{Th}$〈不〉$_{neg}${够健全}}$_{Rn}$，出现〈了〉$_{perf}$[假彩票事件$_{l}$]$_{Re}$，

[e_l]$_A$ 损害〈了〉$_{perf}$[广大彩民的权益]$_P$，{更}$_{ADD}$[e_l]$_A$ 损害〈了〉$_{perf}$[政府的信誉]$_P$，[e_l]$_A${在社会上}$_L$ 造成〈了〉$_{perf}$[严重影响]$_R$]$_{Re}$。

[目前$_i$]$_T$，[杨永明等 3 名涉案人员]$_P$ 被[e_m]$_A$ 刑事拘留，[陕西省和西安市]$_A$ 要求[[公安机关$_m$]$_A$ 加大[案件侦破力度]$_P$，[e_m]$_A$ 尽快破案]$_{Re}$。(中新网)

说明：(i) 从话语结构上看，在标题中，"'宝马'假彩案"是背景性话题，相应地后面的句子"陕西体彩中心主任贾安庆被撤职"是说明部分，用以交代处理的结果。

(ii) 为了简单，我们把"刑事拘留"看作是一个动词性结构而不加分析。

(iii) 从篇章结构上看，这篇报道共有三个段落，第一段只有一句，总提贾安庆被撤职事件；第二段是一个由一连串小句组成的大句，具体介绍贾安庆被撤职的原因；第三段交代其他相关人员的处理结果。这三个段落的衔接很具特色，第二段通过把第一段中报道的前景信息(foreground information)"贾安庆被撤职"名词化为"的"字结构，再作"原因"的定语，即处理为话题性背景信息(background information)，从而自然地把这两个段落联结起来。第三段则用"涉案人员"这种隐性的指示性词语(特别是其中的"案")来回指第二段中的"假彩票事件"，从而在语义上把这两个段落也衔接起来了。当然，上述这些衔接手段都比较隐蔽，并且不易于形式化表示。相对地说，小句和句子之间的衔接就比较紧密，也具有较多的词汇或结构形式手段；比如上例第二段中用了"由于、加之、更"等语篇关联词语。

(3) 标题：{[$e_?$]$_A$ 违规调人}$_{Rn}$[咸阳市人民政府副市长张定会]$_{P/Ex}$[被人大]$_A$ 撤消[职务]$_{Re}$

[新华网]$_A${西安}$_L${6 月 14 日}$_T$ 电([记者边江]$_A$[V])[陕西省咸阳市第四届人民代表大会常务委员会$_i$ 第 26 次会议]$_A${13 日}$_T$ 通过〈了〉$_{perf}$[(关于([e_i]$_A$ 撤消[张定会]$_{Ex}$[咸阳市人民政府副市长]$_{Re}$的)决定]$_{Re}$。

{1999 年前后$_j$}$_T$[张定会]$_{Ex}$ 任[陕西省彬县县委书记]$_{Re}$。

{在此期间$_{j}$}$_{T}$,[工商系统]$_{A}$ 实行[体制改革]$_{P/M/R}$,{[原由地方政府开支的管理的各级工商部门]$_{A}$ 实行[垂直管理]$_{P/M/R}$}$_{k}$,[人们]$_{A}$ 称[e_{k}]$_{P}$ 为["上划"]$_{Re}$。{[彬县工商局$_{l}$]$_{Th}${垂直上划过程中}$_{T}$,存在[([e_{l}]$_{A}$ 严重违规、[e_{l}]$_{A}$ 突击进人[70 多名]$_{Re}$和[领导干部]$_{A}$ 弄虚作假、以权谋私的)问题]$_{Re}$}$_{m}$,[对此$_{m}$]$_{Ta}$[张定会]$_{Th}$负有[直接领导责任]$_{Re}$。{根据《中华人民共和国地方各级人民代表大会和地方各级人民政府组织法》的有关规定}$_{M}$,[咸阳市人大常委会$_{i}$]$_{A}$ 决定[[e_{i}]$_{A}$ 撤消[张定会]$_{Ex}$[副市长职务]$_{Re}$]$_{Re}$。

[与彬县工商局违规调人问题有关的其他数名干部]$_{Th}${也}$_{COR}$受到[相应的党纪政纪处分]$_{Re}$。(新华网)

说明:(i) 相对于后面的"咸阳市人民政府副市长张定会被人大撤消职务","违规调人"的论旨角色可以归入原因。当然,也可以把它处理为原因小句。

(ii)"体制改革"和"垂直管理"对于动词"实行"来说,其论旨角色既像是受事,又像是方式,还有点儿像是结果。为了周全,一并以析取的方式标注出来。

(iii)"称为"可以看作是"称之为"的省略形式,这个被省略掉的空语类是受事论元,其先行语是前面的整个句子。

(iv)"对此"的"此"回指前面的一个大句"彬县工商局垂直上划过程中,存在严重违规、突击进人 70 多名和领导干部弄虚作假、以权谋私的问题"。

(v) 从篇章结构上看,这篇报道共有三个段落,第一段只有一句,总提咸阳市人大撤消张定会副市长职务;第二段由四个句子组成,具体介绍张定会被撤职的原因;第三段交代其他相关人员的处理结果。这三个段落的衔接都缺少显性的表层结构上的标志,但是符合撤职类报道的章法:先总提撤职事件,次述撤职原因,后说相关处理。

(4) 标题:[两名全国人大代表$_{i}$]$_{P}${因[e_{i}]$_{A}$ 涉嫌[贿选和收受贿赂]$_{Re}$}$_{Rn}$被[$e_{?}$]$_{A}$ 罢免

[新华网]$_{A}${北京}$_{L}${12月27日。}$_{T}$电([记者赵磊孟娜]$_{A}$[V]){根据十届全国人大常委会第六次会议27日发布的公告}$_{M}$，分别{因[e_{i}]$_{Th}$涉嫌[贿选和收受贿落]$_{Re}$}$_{Rn}$，[来自湖南和四川的陈满生、郦良钟$_{i}$]$_{P}$被[$e_{?}$]$_{A}$终止[全国人大代表资格]$_{Re}$。

{经查证}$_{M}$：[原任湖南省计划生育委员会主任、第十届全国人大代表的陈满生$_{j}$]$_{A}${在湖南省十届人大一次会议期间}$_{T}$，违反[规定]$_{Re}$，[e_{j}]$_{A}$利用[职务上的便利]$_{Re}$，{通过请客、送礼}$_{M}$，[e_{j}]$_{A}$组织和动员[本机关及本系统的干部]$_{P\&A}$，{为自己$_{j}$当选全国人大代表}$_{Ai}$拉[选票]$_{P}$。[湖南省人大常委会$_{k}$]$_{A}${9月28日}$_{T}$通过[决定$_{l}$]$_{P}$，{[e_{k}]$_{A}$罢免[其$_{j}$全国人大代表职务]$_{Re}$}$_{l}$，[e_{k}]$_{A}${并}$_{ADD}$决定[[e_{k}]$_{A}$提请[全国人大常委会代表资格审查委员会]$_{P\&A}$审查[e_{l}]$_{P}$]$_{Re}$。

[郦良钟$_{m}$]〈原〉$_{past}$任[四川省内江市市长]$_{Re}$，[他$_{m}$]$_{Ex}${因[e_{m}]$_{A}$接受[贿赂]$_{P}$}$_{Rn}$被[e_{n}]$_{A}$依法撤消职务。[四川省十届人大常委会四次会议$_{n}$]{于7月25日}$_{T}$通过[罢免案]$_{P}$，[e_{n}]$_{A}$罢免[郦良钟]$_{Ex}$[十届全国人大代表职务]$_{Re}$。

[公告]$_{A}$还确认〈了〉$_{perf}$[[来自福建的何锦龙和江西的刘和平的全国人大代表资格]$_{Th}$有效]$_{Re}$。{自十届全国人大一次会议以来}$_{T}$，〈已经〉$_{past}$有[余小平、史来贺等7名全国人大代表]$_{Re\&Th}$逝世。

{现在。}$_{T}$，[十届全国人民代表大会]$_{Th}$实有[代表]$_{Re\&Th}$[V][2977人]$_{Re}$。(完)(来源：新华网)

说明：(i) 对应叙述同一个主语的一连串小句，我们尽量为每一个动词性成分标注一个主语，省略的用空语类补充出来。

(ii)"涉嫌"的意思是有跟某件事情有关的嫌疑，其客体论元的论旨角色可以归入系事。"通过决定"中的"决定"跟后面的整个小句同指；它们又跟动词"审查"后面的空语类同指。

(iii) 对于表示存在的"有＋NP＋VP"格式，我们根据 NP 跟"有"和 VP 语义关系，分别标注两种论旨角色。

(iv)"通过罢免案"也可以看作是一个整体充当动词"罢免"的方式论元。

(v) 在“提请全国人大常委会代表资格审查委员会审查”中，兼语式的前段和后段连在一起，兼语成分“全国人大常委会代表资格审查委员会”被标注了两种论旨角色；在“组织和动员本机关及本系统的干部，为自己当选全国人大代表拉选票”中，虽然兼语式的前段和后段被逗号点断，但是兼语成分“本机关及本系统的干部”同样也应该标注两种论旨角色。

(vi) 从篇章结构上看，这篇报道共有五个段落，第一段只有一个大句，总提“陈满生、鄢良钟被终止全国人大代表资格”，并通过原因论元“分别涉嫌贿选和收受贿落”交代原因；第二、三两个段落都是由两个大句组成，分别具体介绍他们的贿选和收受贿落、和被当地人大罢免；第四、五两段交代其他人大代表的情况。这五个段落的衔接都缺少显性的表层结构上的标志，但是符合撤职类报道的章法：先总提撤职事件，次述撤职原因，后说相关事项。

(5) 标题：[田凤山]$_{P}$ 被[$e_?$]$_{A}$ 免去[职务]$_{Re}$[孙文盛]$_{A}$ 出任[国土资源部部长]$_{Re}$

副标题 1：[孙文盛]$_{Ex}$ 出任[国土资源部部长]$_{Re}$

[十届全国人大常委会第五次会议$_{i}$]$_{A}${28 日下午}$_{T}${通过表决}$_{M}$，决定[[e_i]$_{A}$ 任命[孙文盛]$_{Ex}$ 为[国土资源部部长]$_{Re}$]$_{Re}$。

[孙文盛$_{j}$]$_{Ex}$，男，[e_j]$_{Ex}$[V][61 岁]$_{Re}$，[e_j]$_{Ex}$[V][山东威海人]$_{Re}$，[e_j]$_{Ex}$[V][大学学历]$_{Re}$，[e_j]$_{Ex}$[V][工程师]$_{Re}$。[e_j]$_{Ex}$〈曾〉$_{past}$ 任[湖南省委副书记，山西省委副书记、省长]$_{Re}$，{1999 年后}T [e_j]$_{Ex}$ 任[国土资源部副部长(正部长级)]$_{Re}$。

副标题 2：[田凤山]$_{P}${因违纪}$_{Rn}$ 被[e_i]$_{A}$ 免去[国土资源部部长职务]$_{Re}$

[十届全国人大常委会第五次会议$_{i}$]$_{A}${28 日}$_{T}${经表决}$_{M}$ 通过[决定]$_{Re}$，[e_i]$_{A}$ 免去[田凤山的国土资源部部长职务]$_{Re}$；[e_i]$_{A}$ 任命[孙文盛]$_{Ex}$ 为[国土资源部部长]$_{Re}$。[国务院]$_{A}${在提请任免的文件中}$_{L}$ 说，[“[田凤山同志]$_{P}${因有严重违纪问题}$_{Rn}$，〈正在〉$_{prog}$ 调查。”]$_{Re}$

[田凤山$_{j}$]$_{Ex}$，{1940 年 10 月}$_{T}$ 出生于[黑龙江省肇源县]$_{L}$，

[e_j]$_{Ex}$[V][大专文化]$_{Re}$。{1988 年后}$_T$[e_j]$_{Ex}$任[牡丹江市市委书记、黑龙江省副省长、哈尔滨市市委书记、中共黑龙江省省委副书记、省长]$_{Re}$。{2000 年 3 月}$_T$[e_j]$_{Ex}$任[国土资源部部长]$_{Re}$。

[新任国土资源部部长孙文盛$_k$]$_{Ex}$,{1942 年 2 月}$_T$ 生,[e_k]$_{Ex}$[V][山东威海人]$_{Re}$,[e_k]$_{Ex}$[V][大学学历]$_{Re}$,[e_k]$_{Ex}$[V][工程师]$_{Re}$。{1983 年 5 月后}$_T$[e_k]$_{Ex}$任[中共株州市委书记、中共湖南省委副书记等职务]$_{Re}$。{1993 年 9 月起}$_T$[e_k]$_{Ex}$任[中共山西省委副书记、副省长、代省长、省长]$_{Re}$。{1999 年 6 月起}$_T$[e_k]$_{Ex}$任[国土资源部副部长]$_{Re}$。{2003 年 10 月}$_T$[e_k]$_{Ex}$任[党组书记]$_{Re}$。([记者赵磊 韩乔]$_A$[V])

新华网 2003 年 10 月 28 日(中国网)

说明:(i)“免去职务”也可以作为一个动词性结构,不作语义关系的分析。以便跟“免职、去职、离职”等动词相对应。

(ii) 也可以把“十届全国人大常委会第五次会议 28 日下午通过表决”分析为一个小句,于是,后面的“[e_i]决定[e_i]任命[孙文盛]$_{Ex}$为[国土资源部部长]$_{Re}$]$_{Re}$”便要分析为承前省略了主语。

(iii) 为了让每一个动词都构成一个论元结构,所以把承上省略的主体论元都以空语类的形式补充出来了。为了简单,可以认为一般只能作定语的区别词“男”,在新闻等语体中可以作谓语,在词类属性上临时有了动词的功能。

(iv) 在“田凤山同志因有严重违纪问题,正在调查”中,直接以动词的主动形式来表示主语的被动意义;因此,这个主语应该标记为受事。这也许是在新闻语体等特定语境中才有的语法现象。

(v) 这本来是两篇独立的报道,后来被网络编辑按照主题辑合到一起,并加了大标题,但仍然保持原报道的标题(变成了小标题)。其中,第一篇是报道任命事件,共有两段:第一段总说孙文盛被任命,第二段介绍其主要生平事迹。符合一般任命性报道的篇章结构。第二篇是任免报道,共有三段:第一段总说田凤山被免职、孙文盛被任命,第二、三段分别介绍他们的主要生平事迹。也符合一般任命性报道的篇章结构。总的来说,段落之间的衔接,缺少表层结构上显性

的标志。

(6) 标题：[高强]$_{A}$ 澄清：[[[张文康]$_{P}$ 被[$e_{?}$]$_{A}$ 撤职]$_{Th}$ 完全正确]$_{Re}$

孙传炜(北京特派员)

[中国卫生部常务副部长高强]$_{A}$ 说，[“[实践]$_{Th}$ 证明”，[([中国政府]$_{A}$ 撤销[卫生部长张文康]$_{Ex/D}$[职务]$_{Re/P}$ 的)决定]$_{Th}$ 是[“完全正确的”]$_{Re}$]$_{Re}$]$_{Re}$。

[高强$_{i}$]$_{Ex}${近来}$_{T}${因为[e_{i}]$_{A}$〈曾经〉$_{past}$ 公开{为张文康}$_{D}$ 说好话}$_{Rn}$ 而遭到[舆论批评]$_{Re}$。[他$_{i}$]$_{A}${昨天}$_{T}${([e_{i}]$_{A}${在记者会上}$_{L}$ 澄清[自己的有关谈话]$_{P}$)时}$_{L}$ 说：[“其实{[大家]$_{A}$ 如果认真地读一读[我在两次新闻发布会上所披露出来的中国卫生工作存在的各方面问题]$_{P}$}$_{SUP-i}$，{就〈能〉$_{mod}${品味出[张文康工作中存在的失误]$_{Re}$}}$_{CSQ-i}$。”]$_{Re}$

[高强所指的两次新闻发布会]$_{Th/R}$，分别{在4月20日$_{j}$ 和5月30日$_{n}$}$_{T}$ 举行。{在第一场发布会$_{j}$ 上}$_{L}$，[原本担任国务院副秘书长的高强]$_{A}$ 首次{以卫生部常务副部长的身份}$_{M}$ 公开亮相。[他]$_{A}${当时$_{j}$}$_{T}$ 承认，[[北京疫情]$_{Th}$ 远{比官府数据反映的情况}$_{Re}$ 严重]$_{Re}$。[新华社]$_{A}${在会后不久}$_{T}$ 就宣布，[[中共中央]$_{A}$〈已经〉$_{past}$ 撤除[张文康]$_{Ex}$[在卫生部的党职]$_{Re}$]$_{Re}$。

{不过}$_{VER}$，{在第二场发布会$_{n}$ 上}$_{L}$，[高强$_{i}$]$_{A}${却}$_{VER}${为张文康}$_{D}$ 辩护，[e_{i}]$_{A}$ 指[[张文康]$_{A}$〈没有〉$_{neg}${隐瞒[疫情]$_{Re}$}，[[官府数据]$_{A}$ 少报[疫情]$_{P}$]$_{Th}$ 是[{因为“[当时$_{k/?}$ 的信息渠道]$_{Th}$〈不〉$_{neg}${畅通}，〈难以〉$_{mod}${掌握到[准确的数字]$_{Re}$}”}$_{Rn}$ 导致的]$_{Re}$]$_{Re}$。

{“{前几天}$_{T}$，[我$_{i}$]$_{A}${还}$_{ADD}$ 看望〈了〉$_{perf}$[张文康先生$_{l}$]$_{P}$，[和他$_{l}$]$_{D}${就今后加强中国的公共卫生建设问题}$_{Re}$ 进行〈了〉$_{perf}$[深入的探讨]$_{P}$。[他$_{l}$]$_{A}${对我们$_{i+?}$ 的工作}$_{Re}$ 提出〈了〉$_{perf}$[很多很好的建议]$_{R}$。”}$_{q}$[他$_{i}$]$_{A}${在有关场合}$_{L}$ 透露[e_{q}]$_{Re}$。

[高强$_{i}$]$_{A}${当时$_{n}$}$_{T}${还}$_{ADD}$ 表示[[e_{i}]$_{Se}$〈不〉$_{neg}${明白[{为什么}$_{Rn}$[大家]$_{Se}$[对写信给媒体揭露真相的退休老军医蒋彦永]$_{D}$

这么感兴趣]$_{Re}$}]$_{Re}$。

[中国媒体]$_{A}${后来$_{n}$}$_{T}$ 罕见地出现[([e$_{?}$]$_{A}$ 间接甚至直接反驳[这位在职高官言论]$_{P}$ 的)连串报道]$_{Re}$。[《财经时报》]$_{A}$披露,[[经济学者吴敬琏]$_{A}${在给蒋彦永打电话表达敬意时}$_{T}$,批评["{现在}$_{T}$ 有[一种奇谈怪论$_{o}$]$_{Th}$,[e$_{o}$]$_{Se}$ 认为[[说老实话]$_{Th}$反倒有问题,{而}$_{VER}$[封锁消息$_{p}$]$_{Th}$是正常的,[e$_{p}$]$_{Th}$是[对国家和人民]$_{D}$ 负责]$_{Re}$."]$_{Re}$]$_{Re}$。

[最新一期的《经济观察报》]$_{A}${也}$_{COR}$刊登〈了〉$_{perf}$[一篇"瞒报去职,毋庸置疑"的文章$_{q}$]$_{P}$,[e$_{q}$]$_{A}$ 直指[[张文康]$_{Ex}$是{因为隐瞒疫情}$_{Rn}$而下台]$_{Re}$。

《联合早报》(雅虎中国)

说明:(i)"是完全正确的"这种"是……的"结构,可以有两种语义标注方法,第一种像上例所示,认为"完全正确的"这种"的"字结构充当系动词"是"的系事论元;第二种,如第1部分例(10)所示,认为"正确"是谓词核心,"是……的"是强调标记,可以不作语义(论旨角色)关系方面的标注。

(ii)"说好话、感兴趣、有问题"等熟语性结构可以作为一个谓词性成分而不加论旨角色关系分析。"读一读"等动词重叠形式处理为一个整体的谓词。对于形式动词跟其宾语,可以把宾语一律处理为受事;当形式动词直接带光杆的动词作宾语时,也可以把整个述宾结构当作一个动词性成分。

(iii)在"高强……遭到舆论批评"中,"舆论"可以分析为是名词性成分作方式状语,类似于"舆论监督、电话联系"等结构;于是,可以作出如上例所示的语义标注。

(iv)对于用关联词语(特别是配套性关联词语)引导的复句,我们分别用花括号标示各分句的界限,并在括号后标明各分句的语义功能,比如假设性条件(supposed condition,简写为 SUP);结果(consequence,简写为 CSQ)等。还在关联词语下面加着重点。

(v)在"当时的信息渠道不畅通"中,指示词"当时"在上文中没有直接的参照词语;因为这个指示词是直接指示语境中的张文康任卫生部长、非典疫情肆虐而官方少报疫情和隐瞒疫情的那一段时间。

(vi) 动词“透露”的系事论元省略了，其语义就是其前面的直接引语。表示强调的“是”也可以作为其后的动词性成分的一部分而加着重点。

(vii)“为什么”中的疑问代词“什么”的所指是个疑问的、不确定的信息，没有先行语可供回指；因此，不必标注同指关系。在“他对我们的工作……”中，“我们”指的是“我”，只是由于不能或不便、不宜用个人口吻说话，[1]才用这种复数性的人称代词来表示委婉语气。对此，我们用“我”的同指下标加上问号来表示这个代词的所指涉及说话人，并且可能还包括其他相关的人员。

(viii) 这篇报道内容复杂，篇章组织也相应复杂。全文有八个段落，第一段似乎不像是话题句，但包含了标题所透露的主题：“高强澄清：张文康被撤职完全正确”。第二段在首句中通过原因论元“曾经公开为张文康说好话”，交代高强需要澄清的原因。第三至六段具体介绍高强说了些什么和如何为张文康说好话。第七至八段介绍其他相关情况。

(7) 标题：[普京$_i$]$_A$ 签署[命令]$_P$[e$_i$]$_A$ 任免[俄军总参谋长等强力部门高官]$_{Re}$

[中新网]$_A${7 月 19 日}$_T$ 电{据俄通社—塔斯社消息}$_M$，[俄罗斯总统弗拉季米尔·普京$_i$]$_A${星期一}$_T$ 签署[命令]$_P$[e$_i$]$_A$ 任免〈了〉$_{perf}$[一系列俄强力部门高官]$_{Re}$。

[克里姆林宫新闻处]$_A${19 日$_n$}$_T$ 宣布，[{据俄总统令}$_M$，[尤里·巴鲁耶夫斯基$_j$]$_{Ex}${星期五}$_T$ 被[e$_i$]$_A$ 任命为[俄武装力量总参谋长兼俄国防部第一副部长]$_{Re}$，[他$_j$]$_{Ex}$ 接替〈了〉$_{perf}$[原俄军参谋长安纳托利·克瓦什宁大将的职位]$_{Re}$]$_{Re}$。

[克瓦什宁]$_{Ex}${自 1997 年起}$_T$ 任[俄军总参谋长]$_{Re}$。[俄国防部和总参谋部]$_A${在过去 10 年中}$_T$，〈一直〉$_{prog}$〈在〉$_{prog}$ 激烈争夺[军事行动和军队经费掌管权]$_P$。{[俄罗斯国家杜马(议会下院)$_k$]$_A${6 月 11 日}$_T$ 出台[一项法律]$_R$，[e$_k$]$_A$ 授权[国防

① 参考吕叔湘(1980)第 489 页。

部]$_{D\&A}$掌管[武装部队的重要军事行动]$_{Re}$，[克瓦什宁的总参谋部]$_A$主要负责规划[俄罗斯未来的军事进程]$_R$}$_1$。[这$_1$]$_{Th}$就明确〈了〉$_{perf}$[[俄军大权$_m$]$_P$主要[由国防部]$_A$掌管，而〈不〉$_{neg}${是[[总参谋部]$_A$[V][e$_m$]$_P$]$_{Re}$}]$_{Re}$。

{此前$_n$}$_T$[俄《独立报》]$_A$报道称，[[克里姆林宫$_o$]$_A$考虑[将6月21—22日发生的武装分子袭击印古什共和国事件责任]$_P$归咎于[俄武装力量总参谋长克瓦什宁$_p$]$_{L(G)/Re}$，[e$_o$]$_A${并}$_{ADD}$〈可能〉$_{mod}${〈会〉$_{mod}${下达[撤换其$_p$职务的命令]$_P$}}]$_{Re}$。[俄罗斯国防部消息人士]$_A$指出，{因[印古什共和国遇袭事件]$_{Cau}$造成[[大量人员$_q$]$_{Ex}$伤亡]$_R$，[其中$_q$大部分]$_{Th}$是[俄军和其他强力部门的军官及工作人员]$_{Re}$}$_{Rn}$，[俄罗斯总统、武装力量最高司令弗拉季米尔·普京$_i$]$_A$〈已经〉$_{past}$决定[[e$_i$]$_A$撤换[克瓦什宁 r]$_{P/Ex}$]$_{Re}$。{据悉}$_M$，[克瓦什宁 r]$_{P/Ex}$〈已经〉$_{past}$被[e$_i$]$_A$安排到[一个新的职位]$_{L(G)}$，[他$_r$]$_{P/Ex}$〈将〉$_{fut}$被[e$_i$]$_A$任命为[俄安全委员会副秘书]$_{Re}$。

[分析家]$_{Se}$认为，[[[PRO]$_A$指责[克瓦什宁]$_{P\&Ex}$[为印古什遇袭事件]$_{Re}$负责]$_{Th}$是有[一定依据]$_{Re}$的]$_{Re}$。[俄罗斯国家杜马安全委员会成员杰纳基·库特科夫]$_A$表示，[[武装分子$_s$]$_A$成功地{在印古什共和国境内}$_L$发动〈了〉$_{perf}$[袭击破坏行动]$_R$，{并且}$_{ADD}$[e$_s$]$_A$造成[[大量的人员]$_{Ex}$伤亡]$_R$，[颠覆破坏活动]$_{Th}${目前}$_T$〈已经〉$_{past}$成为[俄总参谋部情报总局的"权力"]$_{Re}$]$_{Re}$。

[现年57岁的巴鲁耶夫斯基上将$_j$]$_A${此前$_n$}$_T$〈曾〉$_{past}$担任[俄军第一副总参谋长]$_{Re}$，[他$_j$]$_{Th}$是[俄美削减战略进攻性武器条约$_t$起草官之一]$_{Re}$，[这份文件$_t$]$_P$〈已经〉$_{past}${在({2002年}$_T$[美国总统布什]$_A$访问[莫斯科]$_P$)期间}$_T$签署。

{与此同时$_n$}$_T$，[维切斯拉夫·基霍米洛夫]$_{P/Ex}$被[e$_l$]$_A$解除[俄内务部内卫部队司令的职务]$_{Re}$，[米哈依尔·拉布涅夫]$_{P/Ex}$被[e$_l$]$_A$解除[俄内卫部队北高加索军区司令的职务]$_{Re}$，[安纳托利·叶什科夫]$_{P/Ex}$被[e$_l$]$_A$解除[俄联邦安全局副局长的职务]$_{Re}$。

{据悉}$_{M}$，{[弗拉季米尔·布尔德列夫]$_{P/Ex}$被[e$_{l}$]$_{A}$ 任命为[俄军伏尔加河沿岸—乌拉尔军区司令]$_{Re}$，{同时$_{n}$}$_{T}$ 被[e$_{l}$]$_{A}$ 免去[俄北高加索军区司令的职务]$_{Re}$，[亚历山大·别罗乌索夫]$_{P/Ex}$被[e$_{l}$]$_{A}$ 免去[俄北高加索军区紧急情况副司令]$_{Re}$，{同时$_{n}$}$_{T}$ 被[e$_{l}$]$_{A}$ 任命为[俄国防部副部长]$_{Re}$}$_{u}$。

{此外$_{u}$}$_{TEM}$，[原伏尔加河沿岸—乌拉尔军区司令亚历山大·巴拉诺夫]$_{P/Ex}$被[e$_{l}$]$_{A}$ 任命为[俄北高加索军区司令]$_{Re}$。

[俄武装力量]$_{Th}${除正规军外}$_{Re}$，{还}$_{ADD}$包括[边防、内务、安全、政府通讯、民防和铁道部队]$_{Re}$。[武装力量]$_{P}$[由议会、总统、政府]$_{A}${按照宪法规定的权限}$_{M}$ 共同指挥，[总统]$_{Th}$是[武装力量的最高统帅]$_{Re}$；[军队$_{v}$ 的任务]$_{Th}$是[[e$_{v}$]A 抵抗[外来侵略]$_{P}$ 和[e$_{v}$]$_{A}$ 履行[俄罗斯的国际义务]$_{Re}$]$_{Re}$；[军队人数]$_{Th}$〈不〉$_{neg}${〈得〉$_{mod}${超过[全国人口总数的1%]$_{Ra}$}}；实行[义务兵和合同兵两种兵役制度]$_{Re}$，[服役期限]$_{Th}$分别为[[舰队水兵]$_{Th}$[V][2 年]$_{Ra}$]$_{Re}$，[其他士兵]$_{Th}$[V][1 年半]$_{Ra}$，[受过高等教育的]$_{Th}$[V][1 年]$_{Ra}$等]$_{Re}$。(章田/雅龙)

(编辑：龙猫)来源：中国新闻网

说明：(i) 指示词“这”的参照性先行语为前面的一连串小句。

(ii) 动词“指责”的施事论元是一个隐含的成分。

(iii)“造成大量人员伤亡”的省略的主体性论元，既可以理解为其前句的主语“武装分子”，也可以理解为整个前句(主谓结构)“武装分子成功地在印古什共和国境内发动了袭击破坏行动”。关于这种所指波动的情况，详见袁毓林(2002c)。

(iv) 指示词“此外”的参照语是整个前一段中“据悉”之后的一连串句子。

(v) 为了一句之中相关论元在语义角色上的协调和一致，我们把介词“除”后面的论元的论旨角色归入系事。

(vi) 关于同指关系或参照关系的下标，可以以全文(整篇文章)为单位，这样出现于不同段落中的所指相同的成分可以用相同的下标；但是，这样往往会使下标编号过多。也可以以段落为单位，但是不同段落中具有相同所指的成分无法用相同的下标编号，并且涉及

跨段落的回指或参照关系就难以处理了。因此,比较稳妥的办法是下标全文依次统一编号。

(vii) 这篇报道有十个段落,第一段总提普京任免一批军官,接下来几段分别叙述任免的人物、机构、职务,连带介绍被任命者的生平和被免职的原因,非常符合任免类报道的篇章结构的组织惯例,最后一段介绍相关情况:俄军的体制问题。文章虽然比较长,但是章法十分清晰;并且,用了许多篇章关联词语或具有篇章关联作用的词语来衔接句子和段落,比如“此前、与此同时、据悉、此外”,使篇章结构比较紧凑。

4 结 语

通过对新闻段落、简讯和全文三种长度的文本的语义标注,我们发现它们在语义结构和语义表达方面略有不同,表现为:新闻简讯的论元结构中一般只有主体性论元和客体性论元等动词性成分的必有论元,像时间、处所、方式等非必有论元通常不出现。新闻段落由于不是全文,因而其中有关的代词和指示性词语的先行语和参照语不一定在本段落中出现,也就是说代词和指示词的回指和参照关系可能是跨段落的。即使在全文中,仍可能存在指示情景的代词和指示词,它们并没有明确的参照性词语,而是指示由上文所叙述或暗示的某种情景的。也就是说,这种代词和指示词的先行语没有用显性的词汇形式来实现,其语义是隐含在上下文语境中的。对此,语义标注时无法作出标记,给机器的自动处理带来了难以解决的困难。

另外,我们发现:在新闻文本中,句子之间的衔接,词汇和结构等表层形式手段相对丰富,也便于作出形式化的表示和标注;而段落之间的衔接,词汇和结构等表层形式手段相对贫乏,也难于作出形式化的表示和标注。

参考文献

顾　阳 (1994)《论元结构介绍》,《国外语言学》第1期。
吕叔湘 (1980) 主编《现代汉语八百词》,北京:商务印书馆。

孙 斌 (2000)《继承—归纳机制及其在对象系统中和信息提取技术中的应用》,北京大学计算机系博士学位论文。

徐烈炯 (1998)《生成语法理论》,上海:外语教学与研究出版社。

徐烈炯 (1995)《语义学》(修订本),北京:语文出版社。

袁毓林 (1998)《汉语动词的配价研究》,南昌:江西教育出版社。

袁毓林 (2002a)《信息抽取的语义知识资源研究》,《中文信息学报》第5期。

袁毓林 (2002b)《论元角色的层级关系和语义特征》,《世界汉语教学》第3期。

袁毓林 (2002c)《名词代表动词短语和代词所指的波动》,《中国语文》第2期。

袁毓林 (2003)《走向多层面互动的汉语研究》,《语言科学》第6期。

袁毓林 (2004)《基于论元结构的语义标注的体系和规范》,手稿。

袁毓林 (2005a)《用动词的论元结构跟事件模板相匹配》,《中文信息学报》第5期。

袁毓林 (2005b)《用逻辑和篇章知识来约束模板匹配》,《中文信息学报》第5期。

朱德熙 (1978)《"的"字结构和判断句》,《中国语文》第1、2期;收入朱德熙(1980)《现代汉语语法研究》,北京:商务印书馆。

朱德熙 (1983)《自指和转指——汉语名词化标记"的、者、所、之"的语法功能和语义功能》,《方言》第1期;收入朱德熙(1990)《语法丛稿》,上海:上海教育出版社。

Leech, Geoffrey (1981) *Semantics: The Study of Meaning*. Penguin Books.《语义学》. 李瑞华等译,上海:上海外语教育出版社,1987年。

Lyons, John (1977) *Semantics*, Vol. 1 & 2. Cambridge: Cambridge University Press.

Pan Haihua (2001a) Focus and Scope Interaction in the Interpretation of *Bu*-Sentences in Mandarin Chinese, Lectures in Peking University, June 5, 2001.

Pan Haihua (2001b) Centering Theory and Focus Tracking in Discourse, Lectures in Peking University, June 12, 2001.

2004年7月初稿,2004年9月改定

四、专题研究和个案分析

容器隐喻和套件隐喻及相关的语法现象

——词语同现限制的认知解释和计算分析

本文以"满"、"全"等词语在用法和意义上的差别为例，说明怎样从隐喻的角度分析语言表达中词语的同现限制问题，建议把隐喻分析提升到意象图式的抽象水平，藉此把语言的认知解释转换成算法规则和形式表示，从而实现认知和计算的统一。文章分为八个部分：§1讨论"满＋NP"和"全＋NP"在形式和意义上的不平行性，§2用容器隐喻来分析跟"满"相关的表达，用套件隐喻来分析跟"全"相关的表达；§3用意象图式的概念来讨论不同的容器隐喻表达的心理表征问题，§4用容器隐喻和套件隐喻的中和来解释"满"、"全"跟"一"的交替关系；§5用配偶图式和平分图式的对立，来解释古代汉语中"双"和"两"在意义和用法上的差别；§6讨论隐喻和意象图式的跨平面性(词法—句法)、超范畴性(形容词—动词—副词)和超语言性(汉语—英语)，§7讨论怎样从隐喻表达的图式解剖走向其语义的形式表示和计算分析，§8举例说明认知解释的概括性。

1 "满＋NP"和"全＋NP"在形式和意义上的不平行性

形容词"满"和"全"都可以修饰名词，表示某种东西遍及名词所指的事物。值得注意的是，在这种由形容词修饰名词构成的偏正结构中，它们有时可以互相替换、有时却不能互相替换。例如：

(1) 满身是汗 ～ 全身是汗

满商场的人 ～ 全商场的人

(2) 满脸是汗 ～ *全脸是汗

*满公司的人 ～ 全公司的人

(3) 满场寂静中，舞台灯光忽然聚集在她身上。不等她开

口，全场已是掌声雷动，经久不息。（储引，339）[①]

既然在这种偏正结构中存在着“满”和“全”不能替换的情况，那么说明“满”和“全”在意义上有一定的差别，并由此造成“满＋NP”和“全＋NP”在意义上也有相当的差别。对此，储泽祥（1996）作了详细的考察，得出结论：“满＋NP”和“全＋NP”既能表示范围、也能表示数量。但是，“满＋NP”极言某范围的事物的数量，重在数量，表示范围是附带性的；“全＋NP”直接总括事物的范围，重在范围。例如：

（4）这满屋子的书才是真正的财富呀，一辈子都受用不尽的。（储引，339）

（5）全公司只有我一个人可以在工作时间看报。（储引，339）

他解释道：在例（4）中，“满屋子”首先是表明“书”的数量，由“书”所占据的空间再来表示“屋子”的范围。在“满＋N＋的＋X”中，“满＋N”需要借助X来表示范围，因此（表示范围）是间接性的、附带的。在例（5）中，“全”不需要借助别的成分，直接确定“公司”的范围（第339页）。

储泽祥（1996）的这种说明，不仅跟我们的感觉几乎相反，而且不易说明例（2）中“满＋NP”和“全＋NP”在用法上的不平行性。比如，既然“全＋NP”重在表示范围，那么为什么“全脸是汗”这种说法是不合格的；既然“满＋N”重在表示数量，那么为什么“满公司的人”这种说法是不合格的。如果这两种格式在表示范围和数量上的确有所偏重的话，那么我们毋宁说：“满＋NP”是重在表示范围的，而“全＋NP”是重在表示数量的。因此，当要强调整个脸部这个范围内都有汗时，可以说“满脸是汗”；而遍及脸部的汗是无法（或不易）用数量来度量的，因此不能说“全脸是汗”。当要强调整个公司的人员数量时，可以说“全公司的人”；而公司作为一种非处所性的机构建制，不便用实在的范围来指陈，因此不能说“满公司的人”。当然，这种解释也不

① “储引，339”指转引自储泽祥（1996）的有出处的例句，下仿此。

是最为妥帖的，跟事实还是隔了一层。

比较下列“满＋NP”和“全＋NP”格式在意义上的差别是十分有趣的。例如：

(7) 满场喝采≈全场喝采　　满场的观众≈全场的观众

(8) 满楼的人≠全楼的人　　满车厢旅客≠全车厢旅客

满世界≠全世界

(9) a. 凡是在您手下工作过的同志，调走后都满世界宣传您的事迹。(电子语料)

b. ＊……调走后都全世界宣传您的事迹。

(10) a. 要说全世界各民族让我挑，我还挑中华民族。(电子语料)

b. ＊要说满世界各民族让我挑……

从意义上看，“满场喝采”跟“全场喝采”差别不大；但是，“满楼的人”跟“全楼的人”的差别就比较明显：前者指当时在楼里的全部人员(不管是住户还是临时在那儿的)，后者则指常住在楼里的全体人员。[①] 不仅“满世界”跟“全世界”的意义不同：前者指遍及某个言谈论域(universe of discourse)中的各处，后者指遍及整个地球上的所有地方。而且，这两个格式中“世界”的意义也不相同：前者指言谈中的某个地方，是口语中一种引申性的意义和用法；后者指地球上的所有地方，是书面上一种基本的意义和用法；并且，从所指范围上看，前者明显小于后者。[②] 更耐人寻味的是，引申义的“世界”只能跟“满”组合、不能跟“全”组合，基本义的“世界”只能跟“全”组合、不能跟“满”组合。凡此种种，都激发我们在“满＋NP”和“全＋NP”的形式和意义的不平行性背后，去寻找更具概括性的分析概念和理论解释。

---

① 参考储泽祥(1996：343)§4.2中的有关说明。

② 储泽祥(1996：340)以“满世界”为例，说有些抽象名词跟“满”结合后范围会被缩小。也就是说，他认为这只是词义的临时变化。我们认为也不排除这种可能：“世界”有地方、某处的意思。可资参照的是，在吴语中尚有这种意义和用法的遗存。例如：

该个物事占世界　这个东西占地儿

拿纸头屑粒弄得一天世界　把纸屑弄得到处都有

2 容器隐喻“满”和套件隐喻“全”

事实上，根据我们的语感，似乎应该有比“范围”和“数量”更好的分析概念和理论模型，来解释“满＋NP”和“全＋NP”在形式和意义上的不平行性。例如：（下面的 a 例均来自电子语料）

（1）a. 人家师傅这已经是满肚子不高兴了，……
b. ＊人家师傅这已经是全肚子不高兴了，……
（2）a. 我也是打那时候过来的，满脑子英雄壮举。
b. ＊我也是打那时候过来的，全脑子英雄壮举。
（3）a. 白度和孙过仁心疼地望着元豹，满桌菜肴几乎一口没舍得吃全尽着元豹了。
b. ＊……全桌菜肴几乎一口没舍得吃……
（4）a. 一九七九年在某美术出版社当管子工期间，曾满大街地纠缠女青年，找模特儿。
b. ＊……曾全大街地纠缠女青年……
（5）a. 这是全北京最僻静的地方，坏人作案都不上这儿来。
b. ＊这是满北京最僻静的地方，……
（6）a. 既然是全民族的事就该全民族出血，你不能光指着我们几个派粮派捐。
b. ＊既然是满民族的事就该满民族出血……
（7）a. 我要大声疾呼，让全社会都来关心你们。
b. ＊我要大声疾呼，让满社会都来关心你们。
（8）a. 全单位的人都觉察到阮琳身上将要发生什么不可思议的奇变。
b. ＊满单位的人都觉察到阮琳身上将要发生什么不可思议的奇变。
（9）a. 一个一直坐在一边就餐看了全过程的汉子对女友说：“今儿算是见着真流氓了。”
b. ＊一个一直坐在一边就餐看了满过程的汉子对女

友说……

为什么“满肚子、满脑子、满桌、满大街”和“全北京、全民族、全社会、全单位、全过程”是合格的表达，而相应的“全肚子、全脑子、全桌、全大街”和“满北京、满民族、满社会、满单位、满过程”是不合格的表达。这显然是没法用“范围”和“数量”之类的概念来加以解释的。

在分析诸如此类的语言现象的过程中，我们逐渐体会到：跟“满”相关的语言表达是以容器隐喻(container metaphor)为基础的。比如，人们把肚子看作是一种承受喜怒哀乐、学问、心思等的容器，于是便有了“满肚子的高兴、怒气、委屈、学问、鬼点子”等表达方式；人们把桌子看作是承载饭碗、菜盘的容器，于是便有了“满桌的饭菜、佳肴、酒水”等表达方式；人们把大街看作是一种承载行人的容器，于是便有了“满大街的行人、商贩、特务、小偷、骗子”等表达方式。而北京、民族、社会、单位、过程等稍微抽象一点儿的概念，人们不一定把它们(或它们本身不便被)看作是容器，因此没有“满北京、满民族、满社会、满单位、满过程”一类表达方式。跟“全”相关的语言表达是以套件隐喻(suite metaphor)为基础的，比如：人们把北京看作一个由东城、西城、宣武、朝阳、海淀、丰台、延庆、平谷、怀柔等区县(类似一组部件)构成的一个行政单位(类似一个套件)，于是便有了“全北京的工厂、学校、医院、商店”一类表达方式；人们把民族、社会、单位看成是由更小的单位构成的套件，于是便有了“全民族的力量、全社会的积极性、全单位的职工”等表达方式。人们把事物运动的过程看作是由开始、发展、结束等步骤构成的，于是便有了“生长发育/生命形成/宇宙创生/独立建国/制作动画片/解决问题的全过程”等表达方式。而肚子、脑子、桌子、大街等事物，人们不一定把它们(或它们本身不便被)看作是套件，因此没有“全肚子、全脑子、全桌、全大街”一类表达方式。这就在一定程度上印证了 Haiman(1985)和 Geeraers(1990)等认知语言学著作关于意义的一个基本的观念：语义不是基于客观的真值条件，语义结构也不能简单地化解为真值条件的配列，它并非对应于客观的外在世界，而是对应于非客观的投射世界(projected world)，并与其中约定俗成的概念结构(conceptual structure)

直接关联。[①]

利用容器隐喻、套件隐喻等分析概念，§1中"满＋NP"和"全＋NP"的不平行性就可以得到充分的解释。比如，在人们的观念中，"人"跟"公司"这种抽象的机构难以形成容物（content）跟容器的关系；因此，"满公司"这种表达方式是不可接受的。在人们的观念中，"脸"这种人体部件一般不再分解为几个更小的部件，即它不是套件；因此，"全脸"这种表达方式是不可接受的。同样，在人们的观念中，商场既可以看作是一种承载着店员、顾客这两种人员的容器，又可以看作是一种由店员、顾客双方构成的套件；因此，"满商场（的人）、全商场（的人）"都是可以接受的。据此，为什么在意义上"满场喝采≈全场喝采、满场的观众≈全场的观众"，也可以作出合理的解释：剧场等既可以从整体上看作是承载观众的容器，又可以从构成上看作是由一定数量的座位组成的套件。在这种情况下，具体用"满"还是"全"，全凭说话人更想突出剧场的容器性质还是突出剧场的套件性质。至于"满世界"中的"世界"为什么只能是语义范围较小的引申义、而不能是语义范围较大的基本义，这似乎可以从现实世界知识（world knowledge）对语言表达的制约上作出解释：较小的空间更容易形成容器隐喻，充满这个较小空间具有较大的现实可能性；较大的空间不容易形成容器隐喻，充满这个较大空间具有较小的现实可能性。比如，就"满世界宣传/乱跑/都是商人"等表达而言，引申义的"世界"比基本义的"世界"更可能实现，也更容易被人们理解和接受。

值得注意的是，有时候单独的"满＋NP"或"全＋NP"是合格的，但是当这两种形式后面加上相同的中心语、从而形成"满/全＋NP＋（的）X"格式时，却只有其中的一种格式是合格的。例如：（下面的a例均来自电子语料）

（10）a. 一个穿长衣的小猴打着锣，脖子上拴着绳满场转圈。

b. ？一个穿长衣的小猴打着锣，脖子上拴着绳全场

① 详见张敏（1998），第5—6页；前言，第1页。

转圈。

(11) a. 他……又满身上下摸兜，……。

b. ？他……又全身上下摸兜，……。

(12) a. 丁小鲁几乎全身裸露在雨中，……。

b. ＊丁小鲁几乎满身裸露在雨中，……。

(13) a. 高烧不退，很快出现了中毒性休克，全身各系统随之接连崩溃。

b. ＊……满身各系统随之接连崩溃。

对此，一种可能的解释是：更能跟“转圈、摸兜”这种行为相配合的是空间性的容器，更能跟“裸露、崩溃”这种行为相配合的是“部分—整体”性的套件。

3　从隐喻分析走向意象图式分析

如果仔细追究，那么我们一定会发现：在包含“满”（即以容器隐喻为基础）的语言表达中，容器在空间结构上具有拓扑可变性（立体、平面等）。例如：（均来自电子语料）

(1) a. 看了池里的满池清水，……。

b. ……大厅里充满胜利的欢呼。

(2) a. 华先生的笔脱手掉在地上，他低头满地爬找。

b. 敞开的窗户吹进来的热风使每张办公桌上都落满灰尘。

(3) a. 一骨碌坐起来满头大汗一脸惊恐，……

b. 对面的山上密密匝匝地布满了塔松，……

(4) a. 满树满挂的果子，都着了色，发出香气……

b. 桃花尚未盛开，蓬散为一伞，只枝枝布满花蕾，……

(5) a. 那时咱们也德高望重了，也大大小小满视野了，……

b. 到时候我们也为你们说好话，不搞满门抄斩。

(6) a. 我……怎么一见这孩子就满心高兴？

b. ……心里充满了忧伤。

c. 小姑娘望着分板,〔内心〕充满幻想地说。

d. 我们是很残忍的,〔心里〕充满了杀机。

(7) a. 文字很俏皮,〔文字里面〕充满了英国式的机智。

b. 一直没出声的冯小刚远远地开口,语调浑厚,〔语调中〕充满深情。

c. 要掀起一个学元豹赶元豹的热潮,让生活充满阳光……

d. 上个月底,〔年龄〕刚满十八岁。

e. 我们会把这一天的日程给您排得满满的。

f. 大千世界,无奇不有,清洁工淘粪工〔名额〕都招不满,……

例(1)中的"池里、大厅里"是比较典型的、下凹的、三维立体容器,例(2)中的"地上、办公桌上"是不太典型的二维平面容器,例(3)中的"头、山上"则是不太典型的、向上凸起的、介于二维和三维之间的容器,例(4)中的"树、枝"是不典型的、准圆柱面容器;例(5)中的"视野、门(家族)"则是一种抽象性的、更具隐喻性质的容器,例(6)(7)方括号中的成分是我们添加进去的,其中"(内)心、心里、文字(里面)、语调(中)、生活、年龄、日程、名额"等显然是更加抽象的、饱含隐喻色彩的容器。

上述容器从立体到平面、从下凹到上凸、从具体到抽象、从比较写实的到非常隐喻性的,变化万端。那么,从心理学的观点出发,我们就得提出如下三个问题:(i)人们在语言交际中使用诸如例(1)—(7)这类容器隐喻时,他们在心理上到底有什么样的表征(psychological representation)?(ii)他们的心理上是否真的显现出一个容器的形象来呢?如果是这样,那么这个容器是什么样的(比如:下凹的还是上凸的、三维的还是二维的)?(iii)这种容器是固定不变的,还是会随着不同的语言表达而发生变化的(比如:在例1这类表达中是下凹的、但在例3这类表达中是上凸的)?如果是随着句子而变化的,那么,在例7这类表达中该是什么样的呢?对此,我们信从

Johnson(1987)和 Lakoff(1987)等认知语言学家的见解:[①]这些隐喻的心理基础和表征不是具体的视觉形象,而是抽象的意象(image or imagery)和意象图式(image scheme);比如,容器隐喻之下是一个基于"里—外"关系的抽象的容器图式(container scheme)。根据 Ungerer & Schimid(1996),意象图式是"来源于我们在日常生活中与世界的互动经验的简单而基本的认知结构"。Johnson(1987)说得更具体,他认为:为使我们能具备有意义的、相互联系的经验,并能理解它们及对之进行推理,我们的行为、感觉、知觉活动中一定存在着模式和常规。意象图式正是上述活动中一再出现的模式、形状和规律。意象图式具有下列三个特点:(1)抽象性,它比心理学家所说的心象(mental imagery)更加一般和抽象,跟环境无关;而后者则是一种跟环境相关的较具体的意象,比如看了一个正写的 R 后在大脑中形成不同角度翻转的知觉表征。(2)独立性,它可以超越任何特定的感知方式而独立存在;它主要附丽在感觉运动(sensorimotor)的层面,与我们对空间位置、运动、形状的感受相关;它可以同时是视觉的、听觉的、动觉的和触觉的,是一种空间关系和空间位移的动态类比表征(dynamic analog representation)。(3)完形性,意象图式尽管由可辨识的部分和关系组成,却具有完形的特性(gestalt);是一个内部一致的、有意义的统一体。它是我们获得意义结构的主要方式。有了这些理论上的支持,我们可以回答本段一开始提出的三个问题了。原来,在容器图式这种抽象的认知模式的指导下,外部世界中被我们视为容器的不光包括池子、大厅之类有自然边界的三维实体,还包括被我们感知出边界的地上、树上等实体;总之,凡是有边界或能构想出边界的物理空间都是容器。进一步将这一容器概念映射(mapping)到更为抽象的领域,就形成了各种容器隐喻。比如,视野被概念化为容器,我们的视线界定了视野的边界,它就成为了容器;心理(或心灵)被概念化为容器,各种思想、情感便是盛在这个容器中的内容或容物。不管哪一种容器,它一定具有容器图式的基本结构:

---

① 下述关于意象图式和容器边界等的说明,详见张敏(1998),第 90—102 页;第 103—121 页。

有一个边界，它把相关的空间划分为内部和外部两个部分，从而在人的心理上形成一个容器的构型（configuration）。这种构型具有相当的实在性，所示是意象性的；同时，这种构型又是相当抽象的，适应于不同的在空间上具有拓扑可变性的事物（包括具体的、物质的和抽象的、精神的），所以是图式性的。根据 Anderson（1990：133），图式是一种从特定事例（specific instance）到关于范畴等概括表达的抽象，图式表达（schema representation）可以反映事物的特征构型。因此，把容器隐喻分析抽象为意象图式分析具有充分的认知心理学上的根据。

4 浑然图式“一”：容器隐喻和套件隐喻的中和

上文从容器隐喻表达中容器在空间结构方面的可变性上，引出结论：容器隐喻分析必须上升到更为抽象的意象图式水平，才能更有解释力。这种意象图式概念，对于分析套件隐喻表达就显得更为迫切。比如，人们在使用“全身、全场、全车厢、全单位、全北京、全中国、全民族、全世界、全社会、全过程”等套件隐喻表达式时，不可能真的在心理上形成如下形象：由躯体、四肢和脑袋等构成的人体套件，由一排排座位构成的剧场或车厢套件，由一个个具体部门构成的社会套件，由事物发展的一个个阶段构成的进程套件；而只能是更具概括性的意象图式，这种套件图式由一个整体和若干个部分、一个体现各部分如何构成整体的构型组成。[①] 比如，对于人体套件来说，这种构型就是典型的人体外观：包含五官的头在上、四肢在人体对称的两侧，总之各部分之间在物理上是相连接的；对于社会这种套件来说，这种构型就是一种层级关系（hierarchical relation）：较小的单位逐层构成较大的单位，呈现出一种金字塔形，总之各部分是按照一种抽象的关系而联结成整体的。

① 下述关于意象图式和容器边界等的说明，详见张敏（1998），第 90—102 页；第 103—121 页。

有了意象图式这种抽象的概念，就可以说明在实际的语言使用中容器隐喻和套件隐喻中和化（neutralization）的现象。例如：

(1) 满身是血 ～ 全身是血 ～ 一身是血 ～ 浑身是血
(2) 满身的汗 ～ 全身的汗 ～ 一身的汗 ～ 浑身的汗
(3) 满车厢人 ～ 全车厢人 ～ 一车厢人 ～ 整车厢人
(4) 一着不慎，全/满盘皆输 ～ ？一盘皆输
(5) 满腔热心 ～ 一腔热心　满腔热忱 ～ 一腔热忱
满腔怒火 ～ 一腔怒火
(6) 全心全意 ～ 一心一意

从上述例子可以看出，当容器隐喻和套件隐喻都突出整体性、忽略构成上的细节，即不追究到底是有边界的构型、还是有“部分—整体”这种构型时；可以用“一”来代替“满”和“全”，于是，原来有一定的对立性的容器隐喻和套件隐喻便中和化为更为抽象的浑然图式。至于例(4)中的“满/全盘皆输”不能说成“一盘皆输”，主要的原因是为了跟前面的“一着不慎”中的“一”避免重复。这种用“一”代替“满、全”的浑然意象图式表达，在真实文本中也是极为常见的。例如：（下面的a例均采自电子语料）

(7) a. 一屋人开怀大笑，连于观、杨重也忍不住笑了。
b. 满/全屋人开怀大笑，……
(8) a. 看完稿子已是一身大汗，……
b. 看完稿子已是满/全身大汗，……
(9) a. 少妇一抬手把桌上的杯子扫到地上，接着把一托盘茶杯挨个摔在地上。
b. ……接着把满/？全托盘茶杯挨个摔在地上。
(10) a. 审讯的和被审讯的脸都绿了，一脸不耐烦。
b. 审讯的和被审讯的脸都绿了，满脸不耐烦。
(11) a. 只见刘明顺一头大汗地走在人群前边，……
b. 只见刘明顺满头大汗地走在人群前边，……
(12) a. 一个一身素白，白衣白鞋白头发的小脚乡下老太太……

b. 一个全身素白，白衣白鞋白头发的小脚乡下老太太……

(13) 弯月还想说点什么，忽然发现满店的人都注视着她们，有抽鼻的，有挤眼的，有撮嘴的，一屋子不屑的神色。(储引，343)

上文说这种能跟“满、全”交替的“一”表达的是一种浑然一体式的意象图式，有趣的是，我们的古人对“一”的这种意义特点是有很深刻的理解的。比如，东汉许慎在《说文解字》中对“一”的解释是：“惟初太始，道立于一，造分天地，化成万物”(第7页)。[①]《辞海·语词分册》对“一”的解释更为直截了当：“满；全。如：一天星斗。李煜《清平乐》词：‘砌下落梅如雪乱，拂了一身还满’”(第1页)。[②]《现代汉语词典》基本沿袭这种解释：“满；全：～冬|～生|～路平安|～屋子人|～身的汗”(第1471页)。[③]

Chao (1968)把“一、满、全、整、半、几、多、多、多少、许多、好多、好几、很多”等称为数量限定词(quantitative determinative)，认为它们介于特指限定词(如：每(张纸)、各(国的政府)、某(个人)、本(次大会))和数词(如：一、二、三)之间；指出这种词不给出确切的数目，只指出相对的数目或未知的数目(用于问话时)。赵先生对这种数量限定词“一”的读音有一个极好的说明：这种“一”有完整的重音和变调，即是重读而又有平常的变调。跟没有变调、表示真正一个的“一”不同。例如：

(14) a. 只要一块钱　b. 只要(一)块布

(14a)中的“一”是加重音的数词，指真正一块钱；(14b)中的“一”(包括量词“块”)是轻声，可以省掉，意思是随便任何一块布。数量限定词“一”的意思是“满”、“全”、“整”之类，后头不能用个体量词(单位词)、标准(度量衡)量词和动量词，只用临时量词或容器量词；而且往

① 《说文解字》，据中华书局1979年影印本，以下简称《说文》。

② 据上海人民出版社1977年版。

③ 据商务印书馆1996年版修订本。

往加上“的”跟一个名词,当然也可以不加名词。例如:①

(15) a. 一脸的脏　　b. 一屋子的烟

c. 你看你洒的一身的　d. 一路下雨

可见,“一、满、全”等词在意义和用法方面有许多相似性,而这种相似性又可以追溯到这些词的意义结构背后具有相似的隐喻投射(metaphor projection)和意象图式。

5 配偶图式“双”和平分图式“两”

根据上文的讨论,我们可以说:“满”能够激活(activate)容器这种意象图式,“全”能够激活套件这种意象图式。由于汉字基本上是一种表意文字,因而我们自然要追究:这种意象图式在相应的汉字的字形上有没有一定的反映呢?寻找答案的最简单的办法是查《说文》。《说文》对“满”的说解是:“盈溢也,从水㒼声”(第231页)。从字形上似乎看不出一点容器的痕迹。但是,它的同义词“盈”和“溢”倒是极具启发性的。《说文》对会意字“盈”的说解是:“满器也,从皿夃声”(第104页);对形声字“溢”的说解是:“器满也,从水益声”(第236页);对会意字“益”的说解是:“饶也,从水皿,皿溢之意也”(第104页)。显然,从“满”的同义词“盈、溢、益”等的字形上,可以清楚地看出容器这种意象图式在这些词的意义结构中的作用。《说文》对“全”(古文为“仝”)的说解是:“完也,从入从工”(第109页),对形声字“完”的说解是:“全也,从宀元声”(第150页)。从字形上似乎看不出一点套件的痕迹。但是,跟它们意义相关的“齐”倒是具有一定的启发性的。②《说文》对象物字式的象事字“齐”的说解是:“禾麦吐穗

① 详见Chao (1968),p.578,全译本第487页,节译本第260页。

② “齐”的本义是整齐、一致,后来逐步引申出一同和一齐、同等、齐全等意义。例如:(引自《古汉语常用字字典》,商务印书馆,1993年版,第223页)

(1) 夫物之不齐,物之情也。(《孟子·滕文公上》)

(2) 齐唱田中歌。(刘禹锡《插田歌》)

(3) 与天地兮比寿,与日月兮齐光。(屈原《九章·涉江》)

(4) 佳期别在春山里,应是人参五叶齐。(韩翃《送客至潞府》)

上平也，象形”（第143页）。可见，我们的先哲是用他们最熟悉的“禾麦吐穗上平”这种形象，来反映他们对于“齐”的整齐义的意象图式的。这正好印证了Anderson（1990：133）的断言：图式表达感知信息，不同于命题所表达的意义；图式用以对范畴的典型特征进行编码。我们的先哲在替整齐义的qi这个词造字时，[①]用他们最容易想到的、最典型的齐刷刷的稻/麦穗来代表他们心理上关于整齐的意象图式。可见，从词义的意象图式的角度来分析，有助于了解造字意图跟词的本义之间的复杂关系。[②] 粗略地说，字形有的时候通过全部或部分地描摹形象来直接地反映词义的意象图式、并表示词的本义；比如“盈、益、溢”之类，可以叫直接临摹（direct icon）；有的时候通过举例性地描摹形象来间接地反映词义的意象图式、并表示词的本义；比如“齐”之类，可以叫间接临摹（indirect icon）。

下面，我们讨论直接临摹的一对词及其字形。傅力（1996）在王力（1980：248—252）的有关讨论的基础上，经过仔细体会发现：在古代汉语中，“双”和“两”虽然都指数目“二”，但意义和用法却不同。“双”在句中突出化二为一，强调两者的配合。例如：[③]

(1) 凯子晖与弟恭子，并有时誉，洛阳令贾桢见其兄弟，叹曰：“仆以年老，更睹双璧。”（北史·陆凯传）

(2) 客从远方来，遗我双鲤鱼。（蔡邕《饮马长城窟行》）

(3)（孙）权投以双戟。（三国志·吴书·吴主传）

(4) 卢家少妇郁金香，海燕双栖玳瑁梁。（沈佺期《独不见》）

---

① 很抱歉，在这里我们暂且用“齐”的现代音来代表其当时的古音。另外，关于表意字、象物字、象物字式的象事字和形声字等汉字类型方面的概念，参考裘锡圭（1990）第七章：表意字，第110—150页；第八章：形声字，第151—178页。

② 关于字形和词的本义之间的关系，请看裘锡圭（1990）第七章第二节：字形在词义研究上的作用，第142—150页；第八章第七节：声旁跟字义的关系，第175—178页。

③ 对于傅力（1996）的有关叙述，本文稍微作了一些改动，举例也参照王力（1980）和有关辞书作了调整。例（3）转引自《古汉语常用字字典》第265页，例（7）（8）转引自《王力古汉语词典》（中华书局，2000年）第1611页。

(5) 八月蝴蝶来,双飞西园草。(李白《长干行》)

(6) 何日倚虚幌,双照泪痕干。(杜甫《月夜》)

(7) 其禽加于一双,则执一双以将命,委其余。(礼记·少仪)

(8) 我持白璧一双,欲献项王,玉斗一双,欲与亚父。(史记·项羽本纪)

(9) 双兔傍地走,安能辨我是雄雌?(木兰诗)

(10) 得双石于潭上,叩而聆之。(苏轼《石钟山记》)

例(1)—(3)中的"双"强调事物的配合成对,例(4)—(6)中的"双"强调动作、行为的配合,例(7)—(8)中的"双"强调事物的单位是两个一对,例(9)—(10)中的"双"强调事物的数量是成对的两个。"双"的这种意义特点跟其造字本义是吻合不悖的。《说文》对"只"(繁体为"隻")的说解是:"鸟一枚也,从又持隹。持一隹曰隻,二隹曰雙"(第76页),对"双"(繁体为"雙")的说解是:"隹二枚也,从雔又持之"(第79页)。显然,"双"的造字意图是用一只手提两只鸟的形象来表示合二为一、配成一对这种本义。这种字形比较直接地表示了"双"的这种合二为一、配成一对意义特点背后的意象图式——合而成对,可以简称为配偶图式(one-pair scheme)。

而"两"的意义特点是表示自然界、社会上一种平分为二的现象,其最初的用法是表示具有分而为二特点的事物的数量或单位。例如:

(11) 髧彼两髦,实维我仪。(诗经·鄘风·柏舟)

(12) 两造具备。(书经·吕刑)

(13) 易有太极,是生两仪。(易·系辞下)

(14) 我叩其两端而竭焉。(论语·子罕)

(15) 我两靷将绝。(左传·哀公二年)

(16) 五管在上,两髀为胁。(庄子·人间世)

(17) 两涘渚崖之间,不辨牛马。(庄子·秋水)

(18) 之子于归,百两御之。(诗经·召南·鹊巢)

(19) 与其誉尧而非桀,不如两忘而闭其所誉。(庄子·外

物）

(20) 吾欲两用公仲公叔其可乎？（战国策）

例(11)中的“髦”指朝前向两边分梳为二、下垂至眉的长发，例(12)中的“造”通“曹”，指诉讼的双方；例(13)中的“仪”指太初之时浑然一体的元气判分为二，形成天地，化为阴阳的现象；例(14)中的“端”指事物的一头或一方，事物一般有头尾或始末两端；例(15)中的“靷”指缠束在马胸部用来牵引车轴的两条皮带，例(16)中的“髀”指人体两股的外部；例(17)中的“涘”指水边，一条河通常有两条边岸。上面的“两”是数词，作定语修饰名词。例(18)中的“两”是量词，“百两”是数量词组称代中心语“车”；车子最显著的特征是有两个轮子，所以用“两”为单位；直到现在，“两”仍然沿用作为车的单位，只是字形上增益形旁，写作“辆”。例(19)(20)中的“两”修饰动词性成分，表示在某种意义上具有对立性的两种行为。在上面的例子中，受“两”修饰的名词所指的事物都具有分而为二、两相对立的特点。“两”的这种意义和用法特点跟其造字本义也是吻合不悖的。“两”最初写为“㒳”。《说文》对“㒳”的说解是：“再也，从冂阙。”（第157页），在“两”下云：“㒳，平分”（第157页）。《说文》对“冂”的说解是：“邑外谓之郊，郊外谓之野，野外谓之林，林外谓之冂，象远界也”（第110页）。据此，傅力(1996：382)认为“㒳”的字形示意在于平分，表示自然界、社会上一种平分为二的现象。显然，“㒳”的造字意图是用介空两入的字形来表示分而为二、两相对立这种本义。[①] 这种字形比较直接地表示了“两”的分而为二、两相对立这种意义特点背后的意象图式——分而为二，可以简称为平分图式(two-halves scheme)。

对于两个关系密切、经常一起出现的事物，人们可以突出其相互配合的一方面，即把这种现象归入配偶这种意象图式之中，于是可以用“双”来强调这两者之间的配偶成双的关系；也可以突出其相互对立的一方面，即把这种现象归入平分这种意象图式之中，于是可以用

① 承沈培先生告知，《说文》对“两”的说解是不可信的；有人认为在字形上“两”是两个“丙”，“丙”是马屁股的象形。古代常用两匹马拉车，所以用“两”作车的单位。

“两”来强调这两者之间的分而为二的关系。例如：

(21) a. 葛屦五两，冠緌双止。(诗经·齐风·南山)

b. 未知一生当着几量屐。(世说新语·雅量)

(22) a. 愿君坚塞两耳，无听其谈也。(战国策·赵策)

b. 遂坐而下坠，以双足向前，两手反而后揣草根。(徐霞客游记·滇游日记)

(23) 两水夹明镜，双桥落彩虹。(李白《秋登宣城谢朓北楼》)

在例(21)中，“緌”(帽子上的飘带)用“双”来形容，大概是为了突出其相互配对的特征。鞋类后世通常用“双”作量词，这里的“屦(草鞋)、屐”却用“两(量)”；这是为了突出其分而为二的特征，还是当时的“两”只是如王力(1980：251)所言强调“天然成双的事物”(即不是傅力(1996：382)所言“侧重平分”)，还是因为“双”在先秦时代刚刚出现、还不成熟，[①]这个问题还需要作进一步的研究。至于例(22)中的“耳、手、足”用“两”还是“双”，似乎有点儿随意，这说明“两”和“双”在一定的语境中是可以中和化的。例(23)中对举的“两”和“双”似乎并非随意所为，而是为了塑造特定的文学形象：绕宣城的句溪和宛溪两条河流，相对而流、相互辉映，宛如明镜；宛溪上的凤凰桥和济川桥，上下配合，犹如一对彩虹横跨溪上。这样看来，词义的意象图式在微观上对字形的设计有重要的影响，在宏观上对文学形象的塑造也有积极的影响。

“双”突出配合成偶，强调合作、合并；“两”突出平分对立，强调对抗、分裂。这种意象图式特点在一些流传至今的成语中得到鲜明的反映。例如：

(24) 成双成对～两两相对　比翼双飞～势不两立　名利双收～人财两空

推广开来，汉语量词“双、对、套、副”等跟名词性成分的搭配限

① 王力(1980)指出，在先秦时代“双”字罕见：《诗经》1见，《墨子》1见，《庄子》1见(还是在可疑的《盗跖》篇)，《荀子》中未见(第252页及其注2)。

制，都可以从意象图式的角度进行分析，并能得到合理而充分的解释。

6 隐喻和意象图式的跨平面性、超范畴性和超语言性

从“满、全”和“双、两”等的使用来看，其背后的隐喻和意象图式的作用具有跨平面的特点。它们既可以作为构词语素在词法平面上黏着运用，又可以作为独立的词在句法平面上自由运用；在这些不同的平面上，其背后的隐喻和意象图式是不变的。例如：①

(1) 满额、满分、满怀、满门、满面、满目、满期、满七、满腔、满师、满天、满心、满眼、满意、满员、满月、满载、满嘴、满足、满座，满登登、满堂红；饱满、爆满、充满、丰满、服满、届满、客满、美满、期满、完满、圆满、秩满；满不在乎、满城风雨、满面春风、满目疮痍、满园春色、满载而归、脑满肠肥、心满意足、疮痍满目、春风满面、恶贯满盈、琳琅满目；满打满算、满坑满谷

(2) 全豹、全部、全本、全才、全场、全长、全称、全程、全份、全副、全国、全家、全集、全景、全局、全军、全力、全貌、全面、全民、全能、全年、全盘、全票、全球、全权、全然、全书、全数、全速、全套、全体、全托、全文、全息、全县、全线、全新、全音，全日制；安全、保全、成全、苟全、顾全、健全、两全、齐全、求全、十全、双全、瓦全、完全、万全、圆全、周全，日全食；全力以赴、全神贯注、两全其美、求全责备、十全十美、百科全书、竭尽全力、面目全非、目无全牛、全始全终、全心全意、全知全能、委曲求全、一应俱全、智勇双全

(3) 双边、双鬓、双重、双打、双方、双份、双幅、双杠、双钩、双关、双轨、双簧、双料、双亲、双全、双日、双生、双声、双手、双数、双糖，双喜、双响、双向、双薪、双星，双胞胎、双立人；双管齐

---

① 参考《现代汉语词典》(商务印书馆，1996)和傅兴岭、陈章焕主编(1982)《常用构词字典》(中国人民大学出版社)等辞书，恕不一一具指。

下、一箭双雕、智勇双全、举世无双

(4) 两半、两边、两便、两鬓、两侧、两抵、两地、两端、两广、两汉、两湖、两极、两江、两可、两肋、两立、两利、两免、两面、两难、两旁、两栖、两讫、两全、两手、两头、两厢、两性、两样、两翼、两造、两者,两重性;两败俱伤、两面三刀、两全其美、两相情愿、两小无猜、两袖清风、进退两难、模棱两可、势不两立、首鼠两端、一刀两断、一举两得、一身两役、依违两可

这些都是构词平面上的例子,至于造句平面上的例子请看上文,这里就不再重复了。

"满、全"和"双、两"还有超范畴性的特点,即具有不同的词类功能,充当不同的句法成分。比如,从上文所举的例子可以看出:"满"有形容词(如:满饭桌)、动词(如:满上这一杯)、副词(如:满不是那么一回事)等用法,从而有作定语(如:满办公室)、谓语(如:水库满了)、补语(如:客厅里挤满了人)、状语(如:屋檐上满挂着冰凌)等句法功能;"全"有形容词(如:全农场)、副词(如:那几箱水果全烂了)等用法,从而有作定语(如:全世界)、谓语(如:配料全了)、补语(如:配全了)、状语(如:这些全是水货)等句法功能;"双"在古代汉语中有数词(如:双凫)、量词(如:玉斗一双)、动词(如:其象无双,国士无双)等用法,从而有作定语(如:双桥)、中心语(如:一双)和谓语核心(如:无双)等句法功能;"两"在古代汉语中也有数词(如:两宫)、量词(如:百两御之、葛屦五两)、动词(如:一时无两)等用法,从而有作定语(如:两军阵前)、中心语(如:百两)和谓语核心(如:无两)等句法功能。

有意思的是,在上述不同的语法单位层级、不同的词类范畴和结构功能的情况下,"满、全"和"双、两"背后的隐喻和意象图式的作用却是始终不变的。更有意思的是,跟容器、套件隐喻和意象图式相关的不仅是"满"和"全",而且还有"深、浅、空、缺、齐、套、副、双、对"和"半"以及上文讨论过的"一"等词语。例如:[①]

---

① 参考《现代汉语词典》(商务印书馆,1996)和傅兴岭、陈章焕主编(1982)《常用构词字典》(中国人民大学出版社)等辞书,恕不一一具指。

(5) a. 这口井很深/浅　那间房子太深/浅了

b. 这本书太深/浅　他们俩感情很深/浅

(6) a. 房子空着没人住　她把抽屉腾空了　操场上空无一人

b. 他的话太空,不解决问题　这篇文章很空,没有什么内容

c. 把前面几排座位空出来　空出一天时间购物和旅游

d. 车厢里空得很

e. 屋里连站脚的空儿也没有　抽空儿去一趟北京图书馆

(7) a. 一班全了,二班还缺两个学生　小王还缺四个学分

b. 这本书缺了两页　她又缺了一次课

(8) a. 五十双手套配齐了　一套《全宋诗》买齐了

一副扑克牌缺了三张　一对鹦鹉飞了一只

(9) a. 深闺、深海、深交、深情、深秋、深山、深思、深夜、深意、深渊

b. 浅海、浅见、浅说、浅滩、浅学

c. 满挡～空挡、满腹～空腹、满怀～空怀、满口～空口、满门～空门、全身～空身、满城～空城、满勤～缺勤、满员～缺员、满月～缺月

d. 全额～缺额、全套～缺门、全勤～缺勤

e. 一百～半百、全豹～半豹、?全饱～半饱、全岛～半岛、全价～半价、一截～半截、全年～一年～半年、全票～半票、全日～半日、一生～半生、满身～全身～一身～浑身～半身、一世～半世、全天～一天～半天、一路～半途、全夜～一夜～整夜～半夜、全音～半音、全影～半影

在(5a)中,把井和房子当作容器来谈论其深浅。(5b)则稍微复杂一点,把书和人体当作容器、把书中的内容和感情当作容物;然后,再来谈论其深浅(即容物对容器的占有程度)。在(6a,b)中,"空"读 kōng,是形容词;在(6c—e)中,"空"读 kòng,是动词和名词(不儿化

时是由形容词直接转指成名词，儿化后是由形容词性成分加上转指标记而形成的名词形式）。从(7)—(8)可以看出，跟“满、全”有同义、反义关系的“缺、齐”都可以用容器、套件隐喻和意象图式来解释。(9)则说明跟“满、全”相关的“深、浅、空、缺、一、浑、整、半”在构词时，依然可以用容器、套件隐喻和意象图式来解释其造词的语义理据(semantic motivation)。

值得注意的是，容器、套件隐喻及其意象图式也可以用来解释英语中的一些现象。比如，像名词 capacity 和动词 fill 等是跟容器隐喻及其意象图式相关的。例如：①

(10) a. The assembly hall was filled to capacity.（大会堂里挤满了人）

b. So many people came that the hall's capacity was exceeded.（来了这么多人，以至于大厅里都装不下了）

c. a seating capacity of 1,000(1,000 个人的座位)

d. a capacity audience(满座的听/观众)

e. breathing (or vital) capacity(肺活量)、capacity tonnage(载重量)

f. a capacity to learn (or for/of learning) languages(学语言的能力)

g. He has a mind of great capacity.（他接受力很强）

h. one's capacity as a leading cadre(领导干部的地位)

i. in the capacity of (以……的资格)

(11) a. John filled a glass with water.（约翰在杯子里装满了水）

b. Sounds of drums and gongs filled the air.（锣鼓声充满了天空）

c. be filled with(装满)

① 举例引自《新英汉词典》(上海译文出版社，1985)等辞书，恕不一一具指。

d. This young cadre fills the office satisfactorily.（这位青年干部非常称职）

在(10a—e)中，名词 capacity 通过容器隐喻，指比较实在的容积、容量；在(10f—i)中，则指更为抽象的能力、智能、乃至职位、资格等；也就是说，在人们的经验结构中，他们把人的能力、智能、职位、资格都看成是一种容器。在(11a—c)中，动词 fill 用以指填满比较实在的容器——杯子、天空；而在(11c)中，则用以指填满比较抽象的容器——职位。在同一种容器意象图式的约束下，动词和相应名词的配合十分和谐。而像 component，compound，compose，construct，assemble 等词是跟套件隐喻和意象图式相关的。例子从略。

上面的讨论正好印证了 Lakoff 等认知语言学家的“隐喻的认知观”的下列三个结论：(1) 隐喻的普遍性：隐喻是语言的常态，是人们在使用语言时无须努力就会自动地冒出来的无意识的东西。(2) 隐喻的系统性：隐喻不是个别地、随意地制造出来的，而是有系统的，可形成某种结构化的隐喻群。(3) 隐喻的概念性：隐喻不光是个语言问题，它更是一种思维方式；思维过程本身就是隐喻性的，我们赖以思考和行动的概念系统大多是以隐喻的方式建构和界定的。[①] 在这里，我们要补充的是：从“满”和“全”及相关的语言表达形式上可以看出，隐喻决定了语言的选择和使用，特别是词语之间的同现限制。

7　隐喻表达的图式解剖和计算分析

不同的隐喻反映人们感知事物和事件时的不同的认知方式，从而构成了关于某种事物和事件的不同的意象。意象（或比之更具体的心象），是一种不在眼前的物体或事件的心理表征。比如，当有人要你回忆童年时代在其中度过大部分时光的房子时，你会在他的要求下产生该房子的心理意象，该意象极像一张心理照片。也就是说，

---

① 详见张敏(1998)，第 90—91 页。

在你心灵的眼睛(mind's eye)中,你可能意识到房子的意象突然排列在你的眼前。但是,实际上你的头脑中并没有照片。显然,意象像照片,却又不是照片。那么意象到底是什么?如何解释它的存在?在抽象的、经验性的心理分析层面上,心理学家有形象编码和概念编码等学说;这些不同的学说都能成功地解释一些现象,也都面临着无法解释某些现象的困境。在具体的、物质的大脑—神经的分析层面上,倒是可以肯定地说:意象是神经活动的独特类型(或独特模式、独特位置)的体验。显然,你心理中房子的意象跟组成意象的神经事件并不是不同的事件;相反,意象恰恰仅是这些神经事件。可是,考虑意象时,我们通常是从心理的角度进行的,而不是从神经的角度(诸如神经元及其定位、点火模式、内在联系、发送器物质的数量等)进行的。也就是说,心理层次和神经层次都是真实的,并且可以独立存在。不仅如此,我们还可以在介于心理和神经之间的认知层次上,用意象图式、激活等抽象的术语(而不是神经标签)来描述神经系统经历的神经事件。所谓心理层次相当于我们的意识或觉知(consciousness or awareness),这就是当你在思考你的"心理"时,你所意指的。神经层次是基于或多或少有关神经系统的活动的文字描述的。如果把神经系统的活动描述得更抽象一点,那么我们就达到了认知层次。我们可能不会意识到我们所有的认知和神经活动,但是这些层次都是描述心理事件的一种方便的方式。[①] 于是,在比较抽象的认知层次上,意象可以抽象为结构化的图式,图式可以分解为结构成分及其构成方式。这样,只要找出隐喻表达的构成成分及其结构关系跟相应图式的构成成分及其结构方式之间的映射关系,就可以用产生式规则(production rule)写出算法化的关于隐喻表达的语义解释规则,从而完成从隐喻表达的认知解释到计算分析的技术转变。

比如,对于容器隐喻来说,其意象图式的结构成分是一个边界,

① 参考 Solso (1979) 第十一章:心象,中译本第 307—331 页。Best (1998)第一章:认知心理学:定义、起源和隐喻,中译本第 6—7 页;第六章第二节:分布表象中的有关概念,中译本第 178—179 页。

它把相关的空间划分为内部和外部两个部分，从而在人的心理上形成一个容器的构型。抓住了这一点，我们就可以给出从容器隐喻表达的句法形式到语义表达的形式化的、并且经过调整后是可以算法化的规则系统。例如：

(1) 满桌子糖果　满屋子武器　满脑子小资情调

(2) 满桌子的糖果　满屋子的武器　满脑子的小资情调

(3) 满桌子是糖果　满屋子是武器　满脑子是小资情调

(4) 桌子上放满了糖果　屋子里堆满了武器　脑子里装满了小资情调

(5) 满大街溜达　满地翻滚　满世界找人借钱

如果忽略一些细节，那么例(1)—(3)这三种句法形式表达的意义是相近的；为了方便，可以把这三种格式合记作 S1：满＋NP_1＋(的/是＋)NP_2。作为约定，我们用'NP'代表 NP 的语义所指(semantic referent)。于是，运用一阶谓词逻辑就可以写出 S1 的如下语义解释规则 R1a：

if：满＋NP_1＋(的/是＋)NP_2；then：

{i. 'NP_1' is-a CONTAINER，'NP_2' is-a CONTENTS；'NP_2' is-in 'NP_1'；

ii. $\exists y$，$\forall x$ [is-in (x, y)]→ x＝'NP_2'，y＝'NP_1'}

其中，is-a(属于)和 is-in(在……上/中)等是用以描述语义的元语言(metalanguage)中的谓词，CONTAINER(容器)和 CONTENTS(容物)等是描述语义的元语言中的概念范畴。如果把语句实例"满桌子(的/是)糖果"代入 R1a，那么可以得出如下的语义表达式 M1a：

"桌子"是容器，"糖果"是容物；"糖果"在"桌子"上；

存在着一张桌子，所有的"糖果"都在这张"桌子"上。

显然，像 R1a 这种语义解释规则过于简略，并不能完全反映"满"的充满意义。为了刻画"满"的充满意义，我们必须引入 SPACE(空间)和 SUB-SPACE(子空间)等元语义范畴。于是，"满"的意义可以解释为："满"激活一个关于容器的意象图式，该容器可以划分为若干子

空间，每个子空间中都有容物。据此，语义解释规则 R1a 可以扩充成 R1b：

if：满＋NP_1＋(的/是＋)NP_2；then：

{i. 'NP_1' is-a CONTAINER，'NP_2' is-a CONTENTS；'NP_2' is-in 'NP_1'；

ii. $\exists y, \forall x$ [is-in (x, y)] → x = 'NP_2'，y = 'NP_1'；

iii. CONTAINER has many SUB-SPACE，i. e.，$y = y_1 + y_2 + \cdots + y_n$；

iv. $\forall y_i$，$\exists x$ [has (y_i, x)] → x = 'NP_2'，y_i('NP_1'，i = 1, 2, …, n}

如果把语句实例"满桌子(的/是)糖果"代入 R1b，那么可以得出如下的语义表达式 M1b：

"桌子"是容器，"糖果"是容物；"糖果"在"桌子"上；

存在着一张桌子，所有的"糖果"都在这张"桌子"上；

"桌子(面)"有许多子空间，"桌子(面)"的每一个子空间中都有"糖果"。

显然，R1b 这种语义解释规则在逻辑上是不协调的。因为当我们从外延上把容器划分为许多子空间时，也得从外延上把容物划分为许多子集；也就是说，不能在使用容器的外延意义的同时使用容物的内涵意义。考虑到这一点，R1b 可以修正为如下的 R1c：

if：满＋NP1＋(的/是＋)NP_2；then：

{i. 'NP_1' is-a CONTAINER，'NP_2' is-a CONTENTS；'NP_2' is-in 'NP_1'；

ii. $\exists y$，$\forall x$ [is-in (x, y)] → x = 'NP_2'，y = 'NP_1'；

iii. CONTAINER has many SUB-SPACE，i. e.，$y = y_1 + y_2 + \cdots + y_n$；

iv. CONTENTS has many SUB-CONTENTS，i. e.，$x = x_1 + x_2 + \cdots + x_n$；

v. $\forall y_i$，$\exists x_i$[has (y_i, x_i)] → $x_i \in$ 'NP_2'，$y_i \in$ 'NP_1'，

$i=1, 2, \cdots, n\}$

从数学的角度看，由集合 X 到集合 Y 的关系 R，可以用序对 (x, y) 来表示，其中 $x \in X, y \in Y$。所有有关系 R 的序对构成一个 R 集。在集合 X 与集合 Y 中各取出一元素排成序对，所有这样的序对构成的集合叫做 X 和 Y 的直积集，记作：$X \times Y = \{(x, y) \mid x \in X, y \in Y\}$。显然，$R$ 集是 X 和 Y 的直积集的一个子集，即 $R \subset X \times Y$。[①] 对于这里的 R1c 的 v 行逻辑式来说，关系 has 集是 x 和 y 的直积集的一个子集。如果把语句实例“满桌子(的/是)糖果”代入 R1c，那么可以得出如下的语义表达式 M1c：

> “桌子”是容器，“糖果”是容物；“糖果”在“桌子”上；
> 存在着一张桌子，所有的“糖果”都在这张“桌子”上；
> “桌子(面)”有许多子空间，“糖果”有许多子集；
> “桌子(面)”的每一个子空间中都有一些“糖果”。

像 R1c 这种规则，在逻辑上还算差强人意。但是，在常识和经验方面，可能会碰到不容易自然地处理的实例。比如，像“糖果”这种离散性的物质，划分子集很容易；像“汗水”等连续性的物质勉强还可以划分，因为在“满身的汗水”中，“脸上的汗水”和“背上的汗水”是可以分开的。但是，对“歌声”等连续性的物质划分子集似乎不太自然。比如，碰到“满剧场的歌声”这样的表达，我们能不能把“歌声”划分成几个子集呢？也许，我们可以说：坐在前排听到的歌声和坐在后排或包厢中听到的歌声是不同的。看来，如果对容物进行子集划分是普遍地可行的，那么就可以保证 R1c 在运用上的普遍适用性。

如果上述办法是可行的，那么推广开来，例(4)这种格式可以记作 S2：NP_1 + V 满了 + NP_2。S2 的语义解释规则 R2 可以表示如下：

> if：NP_1 + V 满了 + NP_2；then：
> {i. ‘NP_1’ is-a CONTAINER，‘NP_2’ is-a CONTENTS；

---

① 详见楼世博等(1985)，第 35—36 页。

'NP_2' is-in 'NP_1';

ii. $\exists y$, $\forall x$ [is-in (x, y)] → x = 'NP_2', y = 'NP_1';

iii. CONTAINER has many SUB-SPACE, i. e. , $y = y_1 + y_2 + \cdots + y_n$;

iv. CONTENTS has many SUB-CONTENTS, i. e. , $x = x_1 + x_2 + \cdots + x_n$;

v. $\forall y_i$, $\exists x_i$[has (y_i, x_i)] → $x_i \in$ 'NP_2', $y_i \in$ 'NP_1', $i = 1, 2, \cdots, n$;

vi. ⟨ $\exists A$, $\exists P$, $\exists L$ ['V'(A, P, L)] → A = ∅, P = 'NP_2', L = 'NP_1' ⟩cause⟨λhas⟩}

其中,cause 是描述语义的元语言中的谓词,A 代表施事论元(在 S2 中是隐含不出现的),P 代表受事论元,L 代表处所论元,'V'代表 V 的语义所指,λhas 代表 v 这一行逻辑式。整个 vi 行逻辑式的意思是:A 在'NP_1''V''NP_2'的行为使得'NP_1'的各处都有一些'NP_2'。如果把语句实例"屋子里堆满了书"代入 R_2,那么可以得出如下的语义表达式 M_2:

……(某人)在桌子上堆书的行为,使得桌子上到处都有一些书。

相应地,例(5)这种格式可以记作 S3:满+NP+VP。S3 的语义解释规则 R3 可以表示如下:

if:满+NP+VP;then:

{i. 'NP' is-a CONTAINER, 'VP' is-a CONTENTS; 'VP' is-in 'NP';

ii. $\exists y$, $\forall x$ [is-in (x, y)] → x = 'VP', y = 'NP';

iii. CONTAINER has many SUB-SPACE, i. e. , $y = y_1 + y_2 + \cdots + y_n$;

iv. CONTENTS has many SUB-CONTENTS, i. e. , $x = x_1 + x_2 + \cdots + x_n$;

v. $\forall y_i$, $\exists x_i$[has (y_i, x_i)] → $x_i \in$ 'NP_2', $y_i \in$ 'NP_1', i = 1, 2, …, n}

如果把语句实例"满大街溜达"代入 R3，那么可以得出如下的语义表达式 M3：

"大街"是容器，"溜达"是容物；"溜达"〔发生〕在"大街"上；

存在着一个"大街"，所有的"溜达"行为都〔发生〕在这个"大街"上；

"大街(上)"有许多子空间，"溜达"行为有许多子集；

"大街(上)"的每一个子空间中，都有一些"溜达"行为〔在那儿发生〕。

对于套件隐喻来说，其意象图式的结构成分是一个整体和若干个部分、一个体现各部分如何构成整体的构型。抓住了这一点，就可以参照上文对容器表达的计算分析，把套件的各部分看作是一个个容器，于是套件就成为一套容器；相应地，在这些容器中的容物也成为一套离散的容物。这样，就可以给出从套件隐喻表达的句法形式到语义表达的形式化的、并且经过调整后是可以算法化的规则系统。例如：

(6) 全身伤痕　全身大汗　全单位职工　全世界人口

(7) 全身的伤痕　全身的大汗　全单位的职工　全世界的人口

(8) 全身是伤痕　全身是大汗　*全单位是职工　*全世界是人口

如果忽略一些细节，那么例(6)—(8)这三种句法形式表达的意义是相近的；为了方便，可以把这三种格式合记作 S4：全＋NP_1＋(的/是＋)NP_2。结合处理容器隐喻表达的办法，就可以写出 S4 的如下语义解释规则 R4：

if：全＋NP_1＋(的/是＋)NP_2；then：

{i. 'NP_1' is-a-set-of CONTAINERS，'NP_2' is-a-set-of CONTENTS；'NP_2' is-in 'NP_1'；

ii. $\exists y, \forall x$ [is-in (x, y)]→ x='NP_2'，y='NP_1'；

iii. CONTAINERS is-a SET consists of many SUB-SET，

i. e. , $y = y_1 + y_2 + \cdots + y_n$;

iv. CONTENTS is-a SET consists of many SUB-SET, i. e. , $x = x_1 + x_2 + \cdots + x_n$;

v. $\forall y_i, \exists x_i [\text{has}(y_i, x_i)] \rightarrow x_i \in$ 'NP$_2$', $y_i \in$ 'NP$_1$', $i = 1, 2, \cdots, n$;

vi. $\lambda(x_1, x_2, \cdots, x_n)$ [is-in (x_1, y_1) & is-in (x_2, y_2) & $\cdots$ & is-in (x_n, y_n)];

vii. $\Sigma_x = x_1 + x_2 + \cdots + x_n$}

如果把语句实例"全单位(的)职工"代入 R4,那么可以得出如下的语义表达式 M4:

"单位"是一套容器,"职工"是一批容物;"职工"在"单位"中;

存在着一个"单位",所有的"职工"都在这个"单位"中;

"单位"有许多子集(即部门),"职工"有许多子集;

"单位"的每一个子集(即部门)中都有一个"职工"的子集;

每一个子单位(即部门)中的职工子集的总和就是"全单位(的)职工"。

虽然 R4 看上去是比较折绕的,但是我们希望它能较好地抓住(catch)"全"的语义特点。

最后应该指出,上述"满"、"全"的语义解释规则充其量只是一种非常粗略的逼近。其中,不仅在逻辑上有许多技术细节需要仔细地推敲和修正;而且,从经验上看,有许多参数还需要在具体的上下文语境中依靠百科知识(encyclopedic knowledge)才能设定。比如,语义解释规则 R1b 中的 SPACE 和 SUB-SPACE,在"满大厅的客人"中应该是平面的地面,在"满大厅的歌声"中则应该是立体的空间。另外,作为 SUB-CONTENTS,'客人'是离散的,可以分处于不同的 SUB-SPACE 中;而'歌声'可能是连续的,不一定能分处于不同的 SUB-SPACE 中。再如,语义解释规则 R2 中的 SET 和 SUB-SET,在"全身的汗水"中应该是人体和人体的各个外表部分(脑袋、躯体和四肢等),在"全校的学生"中则应该是班级的全集和一个个班级这种

子集。这些参数都需要在具体的语境中、结合考虑所研究的计算模型的实际应用领域和对象,来设定并逐步加以调整。

8 认知解释的概括性

在传统的分布描写的基础上,从认知的角度对句子中词语之间的选择限制关系进行分析,可以获得直观性很强的、统一而又简明的解释。比如,储泽祥(1996)注意到:

(i)"满"一般只能跟具体名词结合,通常不能跟抽象名词结合(第340页)。

(ii)"满+N"后可以添加"里、上"等方位词(如:满墙上是标语、满堂厅里是客人);而"全+N"后一般不能添加方位词(第342页)。

(iii)"全+N"中的N必须是可总括范围的名词,常要求N有一个完整的范围,范围不确定的名词不能进入(如:* 全地、* 全墙、* 全脸、* 全扁担)(第340页)。

(iv)"全+N"中的N常常是合体的,可以分出层次或不同部分(如:省、市、军区、课程、工序);而"满+N"中的N常常是独体的,难以分出层次(如:被单、柱子、脸、眼)(第341页)。

(v)可以构成"全+N+各+X"格式,不能构成"满+N+各+X"格式(如:全国各地、全厂各车间、* 满城各处、* 满身各器官)(第341页)。

(vi)少数"全+N"格式有对应的"半+N"格式(如:全票~半票、全身~半身、全心全意~半心半意),"满+N"格式一般没有对应的"半+N"格式(第341页)。

对于上述这些看似琐碎并且没有联系的观察,如果从隐喻表达及其意象图式的角度进行分析,那么可以分别作出这样的解释:

(i)因为"满"的意义背后的概念结构是一种容器隐喻,具体的事物比较容易被人们看成容器、抽象的事物被人想象成容器的难度较大;所以具体名词和抽象名词在跟"满"的结合几率上是不均衡的。但是,只要其所指能被人们想象成一个容器,那么再抽象的名词也能跟"满"组合;比如:"满负荷、满工作量、排满了日程"。

(ii) 因为容器表达“满＋N”在意义上涉及空间，所以可以跟方位词组合；而套件表达“全＋N”在意义上不涉及空间，所以不能跟方位词组合。

(iii) 因为“全＋N”是一种套件隐喻表达，所以要求其中的名词必须有一个完整的范围，以形成一个由部分构成的整体的构型。

(iv) 所以要求其中的名词的所指，必须是一种合体的、分层次的部件—整体结构。

(v) 所以可以有“全＋N＋各＋X”这种总—分式表达。

(vi) 因为套件是由离散性的部件构成的，所以能用“半”来度量；而容器是连续性的空间，“半”跟容器的闭合性的边界是不相容的。比如，一个碗是容器，半个碗就不成其为容器了。但是，不排除在形象性的表达(figurative expression)中，可以在一个相对天然的空间的正中间想象出一条边界，从而创造出“半＋N”这种表达方式。例如：

(1) 满窗新绿～半窗新绿　满头白发～半头白发　满池春水～半池秋水

(2) 满圆的月亮～半圆的月亮　满江红(水)～半江红(水)

(3) 一道残阳铺水中，半江瑟瑟半江红。(白居易《暮江吟》)

(4) As for a pessimist, a half full bottle of water is a half empty bottle of water.

(直译：对于一个悲观主义者来说，半满瓶水是半空瓶水；意译：对于一个悲观主义者来说，满满的半瓶水是空空的半瓶水)

可见，认知分析不仅可以解释正确的观察，还能倒过来校正不完全正确的观察。

鸣谢：本文第五节对于有关汉字的字形和字义分析，得到了同事沈培先生的指正；第七节对于“满＋NP＋…”和“全＋NP＋…”等语言表达的语义解释规则的构造和形式表示，得到了同事詹卫东先

生的指正。谨此一并致以诚挚的谢意。当然，如有什么差错，责任全在作者本人。

参考文献

储泽祥（1996）《“满＋N”与“全＋N”》，《中国语文》第5期，第339—344页。

傅　力（1996）《“双”、“两”释异》，《中国语文》第5期，第382—385页。

楼世博等（1985）《模糊数学》，科学出版社。

苏佩斯（1984）《逻辑导论》，北京：中国社会科学出版社。

王　力（1980）《汉语史稿》中册，中华书局。

裘锡圭（1990）《文字学概要》，商务印书馆。

张　敏（1998）《认知语言学和汉语名词短语》，北京：中国社会科学出版社。

Anderson, R. John (1990) *Cognitive Psychology and Its Implications*, Third Edition, New York: W. H. Freeman and Company.

Best, B. John (1998) *Cognitive Psychology*, Heinle and Heinle Publishers, A Division of International Thomson Publishing Inc.《认知心理学》，黄希庭主译，中国轻工业出版社，2000年。

Chao Yuen Ren (1968) *A Grammar of Spoken Chinese*, University of California Press. 据台湾版，敦煌书局，1981年。丁邦新《中国话的文法》（全译本），香港中文大学出版社，1980年，据刘梦溪主编《中国现代学术经典·赵元任卷》，胡明扬、王启龙编校，河北教育出版社，1996年。吕叔湘《汉语口语语法》（节译本），商务印书馆，1979年。

Geeraers, D. (1990) Editorial Statement. *Cognitive Linguistics*, Vol. 1.

Haiman, John (ed.) (1985) *Iconicity in Syntax*. Amsterdam: John Benjamins.

Johnson, M. (1987) *The Body in the Mind: The Bodily Basis of Meaning, Imagination, and Reason*. Chicago: University of Chicago Press.

Lakoff, G. (1987) *Women, Fire, and Dangerous Things: What Categories Reveal about the Mind*. Chicago & London: University of Chicago Press.

Solso, L. Robert (1979) *Cognitive Psychology*, New York: Harcourt Brace Jovanovich, Inc.《认知心理学》，黄希庭等译，教育科学出版社，1990年。

Ungerer, F. & Schimid, H.-J. (1996) *An Introduction to Cognitive Linguistics*. London & New York: Longman.

2001年10月初稿，2002年8月改定

（删节发表于《中国语文》2004年第3期）

关于分词规范和规范词表的若干意见

本文简单地讨论分词时碰到的判定困难和表示困难，指出分词规范应该尽可能地利用规则来说明分词单位的确定原则。最后，建议规范词表应该设立五个不同的等级，以便不同的用户既可以各取所需，又可以互相折算和对应。

1　分词困难的两种类型

大家都认识到，对现代汉语真实文本进行分词会碰到许多困难。就笔者的体会是，这些困难中有两点比较突出。下面，我们作简单的论列。

1.1　判断上的困难

对于结构类型相同、结构项的语法属性相同的字串，哪些是词、哪些不是，不易断定。例如：

鸡蛋 ～ 野鸡蛋　猪肉 ～ 病猪肉　排球赛 ～ 沙滩排球赛

承认前者是词，大概没有问题；说后者也是词，就会引起争议。但是，为什么跟前者严格同构的后者就不是词了呢？可能不容易说出一个令人信服的理由来。也就是说，很难给出一个严格的词的定义，这使得分词在实际操作上会碰到许多见仁见智的分歧。

1.2　表示上的困难

有时候，我们能正确判断哪些字串是词、哪些不是，并且不同的学者之间都有共识；但是，不易揭示这种判断背后的知识。也就是说，不容易找到和说出判断词的标准，难于把词的定义和判定标准表示得明确无误。例如：

牛肉 ～ 马肉 ～ 驴子肉 ～ 长颈鹿肉

订书机 ～ 运煤车 ～ 违章车牵引车

凭感觉，我们可以断定这里的“牛肉、订书机”肯定是词，“马肉、运煤车”就不像是词，而“驴子肉、长颈鹿肉”和“违章车牵引车”就更不像是词了。问题是，怎样把我们的这种语感表示成明确的语言学知识。

2 分词规范中应该利用规则

《信息处理用现代汉语分词规范》§5.1.2.2中指出：民族名、地名中的“族、省、市……”等应单独切分。但包括“族、省、市……”等只有两个字的民族名、地名，则不予切分。例如：

汉族 ～ 哈萨克(族) 忻县 ～ 正定(县)

专名部分不能单独存在而保持原有意义的地名，不予切分。例如：

牡丹江 横断山

街、路、村镇名称，各大洋和大海一律为分词单位。例如：

长安街 大西洋 地中海

我们觉得，上面这些规定，如果用规则说起来，可能会更简单和明了。比如：

如果“专名十类名”能省略为“专名”，那么单独切分，否则不予切分。

3 规范词表应该建立等级

一个通用的兼顾人机的规范词表应该设立等级类别，以便不同的用户既可以各取所需，又可以互相折算和对应。比如，一级词汇是那些没有争议的词，二级词汇是那些游移于词和词组之间的字串，它们通常由一些能产性强的格式或结合面宽的语素造成。例如：

好吃　好弄　公开化　地下化　游戏机　抽油烟机

三级词汇是那些在现代汉语中并不通用的文言词，他们通常出现在固定的文体或结构中。例如：

讯（如“新华社某月某日讯”）

吝（如“这样的事他也吝而不做”）

四级词汇是那些从语言学上看肯定是词组，但同现概率和出现频率极高的字串。例如：

一个　这种　那些　不同　为什么　百分之百

为了信息处理的方便，可以把它们当作是一个分词单位。也可以模仿语音词和语法词的区分，称之为“工程词”。

而对于一些用于汉字输入的词库，像下面这些跨越语法结构层次常见的字串，也可以作为一个分词单位而收入。例如：

这是　那是　不是　不像　很不　不太　也是　都是

这些字串的出现频率是很高的，作为一个分词单位，为输入提供了便利。为了区别，可以称之为“输入词”。也可以把这种“输入词”作为词库中的五级词汇。

如果有了类似的分级词表，那么不同的用户可以根据需要来规定把哪几级词汇作为分词单位。这也许是达到共享和复用电子词典和语料库等语言资源的一种现实的措施。

1997年2月初稿，3月改定

（发表于《语言文字应用》1997年第4期）

2004年9月修改

中文信息处理中的语言难题问答

本文是笔者1998—1999年为《语言文字应用》主持“中文信息处理中的语言难题征问征答”期间，写的一则缘起、四则答问和收集、整理的两组问题。承张伯江先生夸奖这个栏目为“打开《语言文字应用》时，觉得这是一个亮点”。因此，敝帚自珍，也收入这个文集中。

缘　起

由于工作关系，我们经常跟计算机界从事中文信息处理的人士打交道，也经常被他们问起一些语言事实及其相关的分析方法等方面的问题。其中，有不少问题不仅对中文信息处理工作是重要的，而且对语法理论的更新和分析技术的改进也有重要的启发意义。因此，我们假《语言文字应用》这个宝贵的园地，开设中文信息处理中的难题征集和征答这一栏目，供中文信息处理界的朋友能够尽情地提出自己在工作中碰到的语言难题，同时希望语言学界的人士能够踊跃应答和提出各种解决方案。我们热忱地希望这个栏目能够成为沟通信息学界和语言学界的一座小小的桥梁，通过相互问难和辩驳来加强交流、增进了解，共同为推动中文信息处理事业和汉语语言学的向前发展作出贡献。

1　“时间词＋时间词”的结构歧义

问：短语“今天春节”和“今年春节”在词类序列上都是“时间词＋时间词”，但结构关系不同：前者是主谓结构，后者是偏正结构。有没有适当的语法规则，可以据此把它们区别开来？

（北京大学　金茂兵　问）

答：为了区别“时间词＋时间词”的结构类型，必须对时间词进

行精细的次范畴化分类。大体上说，“时间词＋时间词”序列（记作：T1＋T2）可以构成三种句法结构：(1) 并列结构，如：“过去、现在|星期一、星期二|清明、谷雨”；(2) 主谓结构，如：“明天((不)是)中秋|今天((不)是)星期三”；(3) 偏正结构，如：“去年夏天|本月 18 号”。显然地，这三种结构对进入 T1 和 T2 位置上的时间词的小类有严格的选择限制。粗略地说，并列结构要求 T1 和 T2 位置上的时间词属于同一种句法、语义小类，但语义所指一定不同。主谓结构和偏正结构要求 T1 和 T2 位置上的时间词一定属于不同的句法、语义小类，出现在 T1 位置上的通常是“今天、明年”等相对时间词，它们的所指只有参照说话的时间才能确定；出现在 T2 位置上的通常是“星期一、元旦、18 号”等绝对时间词，它们的所指必须参照上文话语中所涉及的时间，比如“星期一”的所指依赖于某个周。此外，主谓结构要求 T1 和 T2 位置上的时间词的所指在量级上是相同的，比如“今年闰年”都是论年、“今天儿童节”都是论天。偏正结构的 T1 一定在量级上大于 T2，比如“今年冬天”，T1 论年、T2 论季。

（北京大学　袁毓林　答）

2 “NP1＋VP＋的＋NP2”的层次切分

问：“我们学校获奖的学生”和“我们学校选送的学生”在词类序列上都是：NP1＋VP＋的＋NP2，但它们的层次构造很不相同。有没有简明的语法规则，据此能够清楚地分化这种层次歧义？

（清华大学　周明　问）

答：为了分化词类序列 NP1＋VP＋的＋NP2 的层次歧义，必须对名词和动词的次范畴（特别是个别动词和名词的配价能力）进行细致的研究。就有限范围内的例子来说，当其中的 VP 是一价的、NP1 是“学生、教师”等零价名词时，NP1 是整个结构的修饰语、“VP＋的”是 NP2 的修饰语、NP2 是 VP 的配价成分（它们有潜在的主谓关系）；当其中的 NP2 是“说法、消息”等有价名词时，NP1＋VP 有可能是主谓结构，它在语义上充当有价名词 NP2 的配价成分，在结构上通过“的”而成为名词性的定语，例如“我们学校获奖的消息（不可

信)”。当其中的VP是二价的、NP1和NP2都是零价名词时,这两个NP有可能都是VP的配价成分;底层的谓词性结构“我们学校选送学生”通过名词化标记“的”提取宾语而变成偏正结构“我们学校选送的学生”。

(北京大学 袁毓林 答)

3 信息处理能不能抛开主语、宾语等概念?

问:印欧语的主语和宾语都是有形态标记的,汉语没有形态变化,主语和宾语等概念是怎样得出来的?在中文信息处理中,彻底抛开主语、宾语等概念有没有可能?

(中国中文信息学会 董振东 问)

答:我们先来考察下列词组,看看能不能从结构上把它们归并为有限的几组:

1. A+B	2. C+D
a. 小王|知道	i. 造|桥
b. 树叶|黄了	j. 买|菜
c. 价格|不贵	k. 坐|火车
d. 今天|星期一	l. 晒|太阳
e. 衣服|晒干了	m. 吃|馆子
f. 大刀|砍钝了	n. 喜欢|闲聊
g. 什么|都不吃	o. 买了|不老少
h. 前面|是条河	p. 来了|几个朋友

上面这些词组都可以分解为两个直接构成成分,并且这两个直接成分之间有某种结构关系。语感告诉我们,a-h在结构关系上比较相近,i-p在结构关系上比较相近。为了方便,我们可以把a-h一类词组所具有的结构关系叫做主谓关系、并把这类词组叫做主谓结构;当然,我们也可以任意地分别叫它们为X-关系和X-结构。同理,我们可以把i-p一类词组所具有的结构关系叫做述宾关系、并把这类词组叫做述宾结构;当然,我们也可以任意地分别叫它们为Y-关系和

Y-结构。相应地，我们可以把主谓结构的前项叫做主语、后项叫做谓语；当然，我们也可以任意地分别叫它们为 X-1 成分和 X-2 成分。同理，我们可以把述宾结构的前项叫做述语、后项叫做宾语；当然，我们也可以任意地分别叫它们为 Y-1 成分和 Y-2 成分。

可见，主语、宾语只是一种方便的称呼，把它们叫做什么并不重要；重要的是要认识到：句法结构是一种关系结构，其各构成成分是受这种关系控制的关系项。因此，一般的语法著作上把主语、宾语叫做语法功能项，那意思就是它们在某种结构关系中扮演了什么角色。基于上述考量，我相信语法研究或者中文信息处理（特别是句处理）可以抛开主语、宾语等名称；但是，必须有 X-关系和 X-结构、Y-关系和 Y-结构、X-1 成分和 X-2 成分、Y-1 成分和 Y-2 成分等概念，以及相应的简称或代号。

这样说来，主语、宾语等句法成分是从句法结构关系中确定的。比如，主语是主谓结构的前项，宾语是述宾结构的后项。而句法结构关系又可以通过相应的变换式系列来确定。比如，主谓结构的前项和后项之间可以插入“呢”等语气词、插入“是不是”构成问句，后项中的动词可以构成“V 不 V”形式；述宾结构的前项和后项之间可以插入“着、了、过”等时态助词，前项中的动词可以构成“V 不 V”形式，整个述宾结构前可以受“不”等否定词的修饰。可见，主语、宾语等概念不仅是可把握的，而且是不可或缺的。例如：

3. NV＋N	4. VP 的＋NP	5. V＋来/去＋了
出租汽车	我买的梨	拿来了
进口设备	他送的书	送来了
研究方法	她写的诗	寄去了
学习文件	爸烙的饼	买去了

对于这种在显性的语法关系方面有歧义的结构，用主语、宾语、谓语、补语等概念来描写和说明是十分方便的。比如，说例 3 既可以是述宾结构、又可以是偏正结构，其中的名动词 NV 既可以是述语又可以是定语、名词 N 既可以是宾语又可以是中心语；说例 4 既可以是体词性的偏正结构、又可以是主谓结构（中间省去“是”），其中的“VP 的”既可

以是定语又可以是主语、NP既可以是中心语又可以是谓语。

（北京大学　袁毓林　答）

4　句型分析和意义分析

问：一般的语法书上说"我们下午开会 ～ 下午我们开会、衣服妈妈洗了 ～ 妈妈衣服洗了、这间屋子我们堆东西 ～ 我们这间屋子堆东西"等都是主谓谓语句。事实上，这里的大主语和小主语分别是施事、时间、受事、工具等；既然如此，设立主谓谓语句这类句型到底有什么意义？

（北京语言文化大学语言信息处理研究所　孙宏林　问）

答：像主谓句、非主谓句、主谓谓语句等都是句型的名称。所谓句型就是句子的结构类型，具有相同的结构模式的句子归在同一种句型之中；反过来说，属于同一种句型的句子具有相同的结构模式。例如：

A	B	C
行！	（我）就来。	我们开会了。
是我。	（她）比你胖。	妈妈洗衣服呢。
起雾了。	（我）不想去。	这个人太认真。
热死我了。	（你）等一会儿！	衣服洗干净了。

对比上例中的A、C两组，可以看出：从构造方式上看，C组的例子都是主谓结构，其中主语和谓语之间可以有一个句中停顿，在这个停顿处可以加上"呢、吧、啊"等语气词，还可以在主语和谓语之间加上"是不是"来构成问句……，正是这种结构上的共性使我们能判定它们都是主谓句；而A组的例子有相当于C组例子中的谓语这种陈述性成分，但缺少相当于C组例子中的主语这种被陈述的成分，因此称为无主句或非主谓句。B组又跟A组不同，A组的例句都是自足的句子，其中补不出主语来；而B组虽然没有出现主语，但是这个主语在具体语境中是可以明确地补出来的，所以是省略了主语的主谓句。

特别要注意的是，主谓谓语句是主谓句下面的一个小类，其特点是谓语部分本身是一个主谓结构。不管是主谓句还是主谓谓语句，作为句型，它们只反映句子的第一或第二层次上的直接成分之间的结构关系，而不反映这些直接成分之间的语义关系。因此，主谓谓语句的大主语可以是施事、时间、受事、工具等语义格，主谓谓语句的小主语也可以是施事、时间、受事、工具等语义格，甚至主谓句或主谓谓语句的谓语中的动词的宾语也可以是施事、时间、受事、工具等语义格。例如：

	D	E
(1)	我们窗户糊了报纸了	他们汽车盖了雨布了
(2)	窗户我们糊了报纸了	汽车他们盖了雨布了
(3)	我们报纸糊了窗户了	他们雨布盖了汽车了
(4)	报纸我们糊了窗户了	雨布他们盖了汽车了

更明白地说，句法结构的构成成分之间的句法结构关系是一回事，句法结构的构成成分之间的语义结构关系是另一回事，主语、宾语等句法成分跟施事、受事等语义成分之间的配位关系(或论元选择关系，argument selection)则又是一回事。语法研究，不管是纯理论的探索还是面向应用的研究，都需要搞清楚从属于同一个动词的各个论元的同现限制关系和它们在句法结构中的位置和顺序。而要想清楚地描写和说明论元结构，句法结构及其类型又是一种必不可少的参照框架。拿 D、E 这类例子来说，我们可以这样来说明动词“糊”和“盖”的论元选择及其句法配置：它们至少能支配施事、受事、材料三个语义格，并且这三个语义格可以在同一个句法结构中共现，这三个语义格可以分别作主谓谓语句的大主语、小主语和谓语中的动词的宾语……。

可见，句型研究这种对句子结构的类型分析，为研究句法成分之间的语义关系提供了描写框架和理论准备。但是，希望从句型上反映出句法结构的语义关系方面的信息是不合理的，除非在纯粹反映句法关系的句型描写上附加语义信息；比如，说 D、E 中的(1)是施事作大主语、受事作小主语的主谓谓语句，(4)是材料作大主语、施事作

小主语的主谓谓语句。如果是这样，那么已经是在句型分析的基础上进行语义分析了。

（北京大学　袁毓林　答）

附录

问题征答(1)

1. 在“今天有雨”和“今天星期天”中，两个“今天”在词性上是否一致？它们的语法(成分)功能是否一样？所表示的语法意义是否相同？是怎样得出是或否这种结论的？

2. 像“那家公司是去年成立的”一类句子在结构上该如何分析？

3. 对于“大权掌握在总统手里”一类句子，有人说其中的“大权”是受事主语。问题是我们是根据什么标准来作出这种判断的？

4. 一个美国学者问：在中国有没有一种公认的汉语句法结构的描述体系和方法，比如像 S → NP＋VP 等的产生式规则及相应的句法分析树？

5. 汉语兼类词的类型有哪几种？有没有区分兼类词的形式化条件？

6. 表示汉语名词复数的显性和隐性标志有哪些？

7. 当句子中有多个 VP 时，作为中心的 VP 的判定条件是什么？

8. 汉语表示被动意义的词汇、句法、或上下文标志有哪些？

9. 汉语名词短语 NP 到底有哪些类型？

10. 英语的时(tense)及体(aspect)在汉语的译文中有哪些相应的表达方式？

以上问题(1)—(4)由中国中文信息学会董振东先生提供，(5)—(10)由国家语委冯志伟先生提供，并经袁毓林先生归纳整理。欢迎大家针对问题作出简明扼要的回答，对每个问题的讨论请限制在2000字以内。同时，欢迎从事中文信息处理的人士把自己在工作碰到的语言难题整理出来、并尽快寄到编辑部来。让我们携起手来，共

同办好这一栏目。

问题征答(2)

1.《现代汉语词典》对“打老虎、打蚊子、打苍蝇”中的动词“打”没给出明确的释义，也没对“打＋名词”这一结构中的名词作出任何语义限制。如果让计算机生成这类短语，如何去防止其说出“打臭虫、打马蜂、打土鳖”一类组合？

2.《现代汉语词典》对动词“洗”的释义是：“用水或汽油、煤油等去掉物体上面的脏东西：～脸|干～|～衣服。”计算机要想学会“洗”的这个用法，就必须弄明白上述释义中谈及的“物体”可以是哪些物体、有没有什么限制。为什么可以说“洗衣服、洗车、洗钱、洗碗”，但不能说“洗玻璃、洗墙壁、洗书本、洗马路”？

3．“北京烤鸭店”中的“北京烤鸭”，根据语感似乎不能理解为“烤北京鸭”或“北京的烤鸭”，应该怎样分析其结构并给出正确的语义解释？

4．“童子烤鸡店”既可以理解为“烤＋童子鸡＋店”，又可以理解为“童子＋烤＋鸡＋店”。在后一种情况下，“童子”既可以理解为烤鸡的人，又可以理解为店名。对其实际意义应该如何给予解释？

5．对于“我想去买点儿东西，然后回家看书”这个句子，能否确定其中的各个成分(包括各动词)之间的逻辑语义关系、并说明各成分在说话人的大脑中出现的先后顺序？怎样来确定这个句子中有没有省略了什么成分？

6．在自然语言处理中，我们经常要用到语义特征去表达句子的某些特殊的语法意义；为了保证这种表达的可计算性，要求每个语义特征有确切的定义、各个语义特征之间有明确的关系。但是，在研究汉语语法的论文中，人们常常是从不同的角度来选取语义特征的。以动词为例，从动作主体上，有[述人]、[非述人]、[可控]、[非可控]、[自主]、[非自主]等；从时间上，有[完成]、[持续]等，从词汇意义上，有[动作]、[变化]、[位移]、[取得]、[给予]、[制作]、[附着]、[取除]、[破损]、[致使]、[感受]、[状态]等。这些语义特征之间有什么关系？能不能把它们汇成一个有序的集合？

以上问题(1)—(5)由中国社会科学院语言研究所杨国文先生提供,(6)由中国社会科学院语言研究所傅爱平先生提供,并经袁毓林先生归纳整理。欢迎大家针对问题作出简明扼要的回答,对每个问题的讨论请限制在2000字以内。同时,欢迎从事中文信息处理的人士把自己在工作碰到的语言难题整理出来、并尽快寄到编辑部来。让我们携起手来,共同办好这一栏目。

(发表于《语言文字应用》1998年第3、4期)

缓冲式移动通信及其发展方向

——一个语言学家的设计思想

本文从语言学角度，分析语言交际的时空局限和怎样用符号和技术手段来突破这种限制，涉及到文字、书信、电话、电报、录音电话、传真、移动电话（手机）、电子邮件和手机短信等交际工具或手段。特别是建议设计录音手机以推动缓冲式移动通信，最后着重讨论了几种打破手机短信的汉字输入瓶颈的方案。

1 用文字符号突破语言交际的时空限制

人类发明了语言，用以交流思想和协调行动，并组成足以主宰自然界的社会群体。这段历史，最保守的估计也有几十万年。但是，语言作为一种以声音为媒介、诉诸听觉的交际系统，在使用上存在着时间和空间两大限制；即在时间上必须是同时性的听和说的轮替和反馈，在空间上必须是同地性的、让耳朵的听觉够得着的。这种时空限制使得异时、异地的人们无法用语言来进行交际。为了克服这种时空限制，人们又发明了文字来记录稍纵即逝的语言，并形成了跟口头语言不完全相同的书面语言。这种使用文字的历史，最大胆的估计也不会超过一万年。

有了文字，大到祖宗的事迹和先圣的哲语可以垂示后人，远方的民情和边疆的战况可以用文书来呈报朝廷；小到买卖或租赁双方可以签订契约，长辈在临终前可以立下遗嘱，指示子女们该怎样来分割遗产。在民间，书信往来成为一种最有效的不受时空限制的语言交际。当然，这种书信往来式交际的反馈是十分滞后的；并且，传递信件是要付出很大的代价的。比如，中国古代官府耗费大量民脂民膏来修筑驿道，专门用以传递政府公文。可惜的是，这种专门为通讯服务的交通设施，并未惠及寻常百姓，更未带来通讯的社会化和商业

化。

上面所述的这种情况，可以概括地称为用文字符号来突破语言交际的时空限制。

2 用技术手段突破语言交际的时空限制

进入工业革命时代以后，在近代自然科学和声、光、电技术的激励下，电话发明了，使得远在千里之外的人们可以直接对话。相应地，人类的口头语言交际发生了一场深刻的革命。电话在战争中起了十分重要的作用，古人说的"运筹于帷幄之中，决胜于千里之外"得到了真正的实现。这可以看作是用技术的手段来突破口头语言交际的空间限制，但是电话显然未能突破语言交际的同时性限制。随后，无线电技术的进步促使了电报的发明，于是，书信这种人类的书面语言交际样式在速度上发生了革命性的变化。从理论上看，电报倒是继承了文字和书信那种突破语言交际的时空限制的优良传统，并且在速度上又克服了书信的邮递周期长的缺点。但是，电报依赖于编码和译码，无疑把书信交际代价高的缺点大大地放大了。即使是精简到干巴巴的电报式语言，也未能抵销这种高昂的代价。因此，电报从来就没有成为人们的一种普遍使用的交际方式。

随着电磁技术的发展，录音的技术和装置有了长足的进展。于是，把录音设备负载在电话上就成为录音电话。这在一定程度上可以克服语言交际的同时性限制。但是，录音电话无法实现人们口头交际的一个基本的要求，即及时地反馈信息并形成听说的轮替。因此，录音电话在现在也并不是人们普遍使用的通讯设备。传真机的发明，有效地克服了书信往来周期长、电报需要编码和译码的缺点。并且，除了书信之外，传真机利用光电效应还可以把照片、图表、证件、文件等的真迹传送到远方。因此，随着传真的普及，电报业务日趋萎缩。虽然，在双方都拥有传真机的条件下，可以形成快速的书面交际反馈；但是，书面语言不如口语那样简便、灵活，因此电话仍是远距离通讯的最方便和最有效的手段。

3　通信设备的移动化和杂交化

除了同时性限制之外，电话的另一个缺点是：设备固定，不能随身携带。结果，虽然城市中到处矗立着电话亭，但是人们仍然觉得不能得心应手地使用，主要是不能随时随地想用就用。为了对付这种困难，移动电话（手机）就应运而生。在克服了成本高这一因素以后，手机现在使用得十分普遍。并且，手机还可以跟家里的座机连结起来，在家中无人值守时让座机电话转发到手机上。这就充分实现了口语通讯的移动化和便捷化。

跟这种远距离口语交际的技术革命差不多同时，远距离书面语交际也在进行轰轰烈烈的技术革命。随着网络技术的发展，把电话线（或网线）跟个人电脑连在一起，就可以进行电子邮件的交互传送。其传输速度之快使得交互反馈非常迅速和及时。这样，不要说电报，就是传真都受到排挤了。并且，还可以手机电脑化，增加存储和显示功能；于是，就可以把电脑跟手机连结起来，让电脑上的电子邮件在手机上显示出来。这样，不同的通讯设备真正做到了杂交并发挥出功能上的杂交优势。更进一步，直接在手机上输入和收发短信。使得在深夜、开会、上课、看演出等不便使用有声语言的场合，也可以用书面语言来进行语言交际。这就难怪手机短信现在竟成了痴男怨女们谈情说爱的最佳方式，因为它可以快速地表达一些羞于说出口的情感，又可以夸张地措辞，而对方则可以从容地选择应对的方式和词句。

4　手机短信怎样突破输入的瓶颈？

随着手机短信的广泛使用，汉字输入的瓶颈就显得十分突出。怎么办？至少有两种思路，一种是绕着走，回归到手机作为有声口语交际的远程通讯工具这一本来面目上来。比如，在手机上增加耳麦就可以在大庭广众不影响别人地听电话；同样，在手机上增加录音装置则可以把电话留存起来，等到方便的时候再听。也就是说，录音手

机的市场需求是迫切的，应该尽早地推出来。对于关机的用户，通讯总部可以给予保留，等到他开机时再把录音发给他。从理论上讲，录音手机是一种缓冲式的移动通讯工具，可以避开不合适的时机又不影响最终听到电话。当然，这种缓冲式通讯的反馈势必也会滞后，不能形成及时的交互通讯。

另一种是在手机上设置书写功能。但是，现在手机上用拼音输入法书写短信殊为不便，一个键上印了三个拼音字母，按一下这三个字母都出现，再按选择健选择，……最后才从一串同音字中找到你要的那个汉字。挑选标点符号更为不便。这极大地限制了手机作为书写和传递书信的通讯工具。

因此，我们必须设法打破文字输入的瓶颈，通过增加按键、扩大屏幕、改进输入方案、储存常用语句、优化菜单等，在硬件和软件两方面着力。其中有一种彻底的做法是：增加书写板，把手写的短信作为一个图像发出去。显然，这种方式传输的数据量大，手机的造价也会增大，还存在诸多技术上的困难。一种讨巧的办法是，增加一个跟电子词典的接口。这样，可以把电子词典兼用作输入键盘和打字机。如果电子词典中存有大量的词汇和常用语句，并且提供多种检索和输入方法；那么，这种外接方式将是很有市场前景的。还有一种极为朴素和笨拙的做法是回到电码本时代，充分利用常用汉字数量有限的特点，把7000个通用汉字编成数字代码手册。当然，手册上要提供拼音、部首等多种检索方法。这样，任何一个汉字一定在4位数之内得到了编码；并且，越常用的越靠前，因而数字越短。用户边查手册边输入数字，端的不方便，但是十分保险；因为你总能用你最熟悉的方法查到你要的汉字，相应地也就找到了数字代码。诸如此类的方法还有许多，拿出来在现代通讯技术的背景上检讨一下，看看怎样改造和组合，也许可以帮助我们拓宽思路，为打破文字输入、乃至远程通讯的某些瓶颈提供实用的门径。

2004年4月初稿，9月改定

走向多层面互动的汉语研究

本文主要评论汉语研究中关于语法、词汇、语音相互关联的有关研究。首先，指出传统的语言研究语音、词汇、语法各自独立，描写语法学的操作程序又强调只能先分析音位、后分析语法，到了生成语法时代才确立低层面的语音分析和高层面的语法分析应该互动的正确观念。接着，从汉语的结构面貌和汉语语法研究发展史的角度，指出从词法、形态本位向词组、句子本位转移是汉语语法研究的必然趋势；还指出语法分析中的句法、语义、语用三个平面除了要注意区分之外，更应该研究这三个平面上的成分之间的配位关系和相互关联的标记模式，并考察语义、语用因素的语法化过程，从而沟通语言的共时研究和历时研究。然后，着重评述有关汉语语法和语音互动关系的研究，指出国内学者在传统语法和描写语法的框架内，对于语音对语法的制约作用已经作了初步的探索；国外学者在生成语法和生成音系学的背景上，分别探讨了语法结构对于语音(特别是方言中连读变调域的界限)的制约作用和韵律结构(特别是音步组织)对于语法结构的制约作用，建立起了各种颇具理论语言学色彩的理论模型。最后，讨论词库结构和句法操作的互动关系，指出假设在介于词库和句法表达之间的论元结构层面上的有条件的变化，可以免去许多繁复的句法操作。

1 语言研究的三分格局和互动观念

传统的语言研究，基本上是语音、词汇和语法三分天下，呈鼎足而立之势。这种语言研究的三分格局(tripartite paradigm)，直到上世纪初索绪尔(De Saussure)创立了现代语言学，依然没有受到多大的影响。因为在《普通语言学教程》中，语言被看作是符号系统，符号具有声音和意义两重性(duality)，音义结合的符号可以指涉(denote)或代表(represent)外部世界或精神世界中具体的或抽象的事物；于是，语音研究和词汇研究都是语言研究的重要的部门。另外，

索绪尔强调符号之间具有横向的组合关系和纵向的聚合关系；于是，语素组合成词、词组合成短语、短语组合成句子、句法成分之间的形态上的一致关系等词法和句法组合关系，以及形式上的聚合关系等形态问题、语法意义上的聚合关系等语法范畴问题、同功能（分布）的词之间的聚合关系等词类问题，都必须纳入语法的两个部门（词法和句法）中进行研究。后来，在布龙菲尔德（L. Bloomfield，1933）的《语言论》，乃至霍凯特（C. Hockett，1958）的《现代语言学教程》中，也都是顺着语音、词汇（语素或意义）、语法的次序一路讲过来。至于这三个部门之间的关系，好像没有人正面说有或无；反正在实际上一般是各自为政，分而治之的。

到了后布龙菲尔德学派（post-Bloomfieldian school）那儿，情况有了很大的改变。因为，他们强调语言研究的发现程序（discovery procedure）；于是，只能通过对立、互补等分布分析的操作程序，先发现能区别意义的最小的语音单位（即音位），然后再发现最小的有意义的语言单位（即语素），最后发现语素类及其配列方式（即句型）。[①] 按照他们的理论，这些步骤之间不能窜改，否则就是自乱步伐。一旦这种工作假设成了教条，就连极有田野工作经验的语言学家，也只得把它们奉为不可逾越的金科玉律。最有趣的例子是董同龢先生，他在《四个闽南方言》（《历史语言研究所集刊》第三十本，第 729—1042 页，1959 年）中写道：

> 所谓“变调”和“本调”不同是在实际语言中语位和语位相接的时候才显示出来的。以前曾说：“凡单独出现，在轻声字前，在句尾，在主语之末，在表时地的副词之末，在名词或动词系列中各名词或动词之尾的，同是一个调值；在别个字前面而不属上述各种情形的，又另是一个调值。”（《厦门方言的音韵》，《历史语言研究所集刊》第二十九本，第 242 页，1958）那大致是从语法上的地位来分，说起来简单而容易了解。不过，近来有人觉得：在语言分析的过程中，如果在作音位分析的

① 袁毓林（2000b）对此有一个简要的总结，更详细的讨论，请看 Hockett（1942，1947）和 Harris（1946，1951）等。

> 时候就谈语法上的区分，理论上似乎颇有问题。如此，作者以为闽南话"变调"和"本调"的不同未始不可以照他们的另一个说法解释作语音的"接合形态"(junctural feature)的不同。换言之，凡用"变调"的是和后面的成分接合比较紧凑的，凡用"本调"的是和后面的成分接合比较松弛或者后面有停顿。总之，无论怎样去解释，这都是语言分析上，尤其是形态音位学方面，一个很有意思的问题，值得我们多加思索。暂时，我们以记出现象为已足。

对此，李荣(1983：14)作出了非常有启发性的评论：

> 先讲音韵，后讲语法，音韵讲完之后才能讲语法，否则叫自乱步伐，理论上有问题。这只是美国某一派语言学家一时的主张，并非颠扑不破的理论。其实在这里，用变调表示跟后头成分接合得紧，用本调表示跟后头成分接合得松。结合得松紧是从用本调还是用变调推出来的，并无其他标准可以判定。用本调还是用变调，结合得紧还是松，是一件事的两方面，并不是一种解释，从语法地位分才是真正的解释。大家都知道，从一九五八年左右开始，那一派语言学家的主张在美国就失势了，再也不是主流派了。由此可见，迷信一种流行的语言学理论是要上当的。研究语言的人千万别忘了，实践是检验真理的标准，语言比语言学丰富，语言学理论必须建立在语言事实的基础上。

李先生的分析和议论是非常发人深省的。如上所述，后布龙菲尔德学派语言学的目标是：通过执行一组对语料的操作来"发现"语法。每次连续的操作就从语料里进一步地去掉一个步骤。由于一连串言语的物理记录是唯一客观的起点，因而要达到语法描写的层面，就必须遵照以下的次序：(i) 音位学(phonemics)，(ii) 语素学(morphemics)，(iii) 句法学(syntax)，(iv) 话语(discourse)。由于先要从言语流中把组合在一起的音位抽取出来以后才能找出语素，因而在音

位描写中不能引入语素(或句法)的信息。[1] 比如,Hockett (1942: 19)就明确地指出:

> 一定不能有循环论证的现象。我们是为了语法分析才进行音位分析的,因而在分析音位的时候不能有语法部分的任何假设。两者的界线必须划分清楚。

其实,Hockett (1955: 3)已经指出,像 Kenneth Pike 等研究者就坚信:为了理解音位系统,不仅需要知道相关的语法系统的一些事情,而且要积极地利用语法知识来作为音位判断的标准。至于生成语法学者,他们既然不认同后布龙菲尔德学派的经验主义研究理念和发现程序,自然也不会理会这种先音位后语素、句法等操作顺序方面的教条。比如,Chomsky (1957: 59)就毫不含糊地说:

> 语言描写的高平面有赖于低平面获得的结果,这的确是一个事实;可是,反过来也是事实——低平面的描写有赖于高平面上的结果。(中译本第 58 页)

可见,进入生成语法时代,语法和语音这两个平面应该互动(interaction)的观念是牢固地确立了。接下来的任务是,怎样在分析相关的语言现象时落实这一观念,并提出具体的研究步骤和操作方法。

2　语法研究的两个部门和本位意识

根据传统的说法,语法研究可以分为词法(形态)学和句法学两个部门;前者研究语素怎样构成词,后者研究词怎样构成句子。[2] 对于有形态屈折变化的语言来说,像名词的性、数、格和动词的时、体、态,乃至词类范畴等大部分语法信息,都被纳入词法(形态)学之中。

---

① 参考 Newmeyer (1986), p. 7—11;中译本,第 8—13 页。

② 根据 Crystal (1997),按狭义理解(即按语言学的传统涵义和通行的理解),语法指语言结构的一个层面,可独立于音系学和语义学进行研究,通常包含句法学和形态学两个分支。按这一涵义,语法是研究词与构词成分如何组织起来形成句子。广义的语法指一种语言的结构关系的整个系统。于是,除了句法学之外,语法学还包括音系学和语义学。详见中译本第 163—164 页。

词法（形态）学实际上又至少包括构词法和构形法两个子部门，构形法包括了形态变化、语法范畴和词类等语法的主要内容。于是，真正的句法学部分的内容就显得十分贫乏。[①] 比如，Jesperson（1933）共三十六章，其中首章和末章讲导论和回顾，第二～六章讲语音及拼写，第七、十四～二十七、三十一、三十二章讲词类总论、格、人称、代词、性、数、级、时、助动词（will，shall，would，shoud）、式（mood）、分词、不定式，第八～十三、二十八～三十、三十三～三十五章讲三品（the three ranks）及小句作三品（首品、次品、末品）、附加与核心（junction and nexus）及其各种表现、句子结构、动词跟主语与宾语的关系、被动式、谓语、肯定、否定、疑问。讲词法的篇幅远远多于讲句法的，并且像肯定、否定、疑问又是从表达的角度讲的。总体上是从理解的角度来组织语法知识的，中间又穿插从表达的角度来叙述；因此，整个体系显得有点儿凌乱。传统语法如此，到了结构主义语法依然没有太大的改观。比如，根据 Hockett（1954）的说法，在描写语法现象和组织语法知识方面，传统语法学用的是词和词形变化表（word and paradigm）模式（简称 WP 模式），结构语法学用的或者是项目和配列（item and arrangement）模式（简称 IA 模式），或者是项目和变化（item and process）模式（简称 IP 模式）。顾名思义，WP 模式以叙述词形变化来组织语法知识，结果自然是语法学约等于词法（形态）学。IA 模式和 IP 模式本来可以侧重分析句子的构造，但是 Hockett（1954）举的例子主要是怎样描写动词过去式的构成方式。比如，bake＋-ed＝baked，是 take＋-ed＝took 还是 take＋a → o＝took，等等。结果，给人的印象依然是语法学约等于词法（形态）学。当然，结构语法学跟传统语法学相比，在句法分析方面的进步还是有的，只是并不太多而已。比如，提出了结构层次（structural hierarchy）的概念，用以说明和分析 old men and women 一类结构歧义现象。特别是布龙菲尔德之后的学者，还发展出替换、扩展、紧缩乃至

① 详见陈平（1988）。

概率统计等一整套直接成分(immediate constituents)分析方法。[①]后来,Zellig Harris 又提出了变换分析法(transformational analysis)。除此之外,好像说不出太多的内容来了。真正的句法研究的繁荣,那是 Noam Chomsky 创立转换生成语法学以后的事情了。也就是说,五十年代末以后句法学才真正成为语言学的中心。

汉语语法学是在西方语法学说的强烈影响下发展起来的,一开始的语法学模式自然是词本位的语法学。比如,马建忠出版于 1898 年的《马氏文通》,是中国第一部比较系统的语法著作。全书共十章,绪论部分为"字类、句读"正名,第一～九章讲名字、代字、名代之次、静字、动字、状字、介字、连字、助字和叹字,第十章讲起词、语词、止词等句法成分和顿、读、句等句法单位。汉语虽然没有形态变化,但是众多的虚词也足以使得这种以词为中心的语法体系羽毛丰满。黎锦熙 1924 年出版的《新著国语文法》,是我国第一部白话语法。其中提出了句子本位的语法,在当时堪称是一种新潮的观念。全书共二十章,第一章为绪论,讲词、句、词类和句法及其关系等;第二章讲词类的区分和定义,第三章讲单句的六大成分及其图解法,第四章讲实体词的七位,第五章讲主要成分的省略;第六～十一、十五～十八章讲名词、代词、动词、形容词、副词、介词、连词、助词、叹词及相关的句法问题;第十二章讲单句的复成分(即词组作句子成分),第十三章讲附加成分的后附,第十四章讲包孕复句(即小句作句子成分);第十九章讲篇章和修辞,第二十章讲标点符号。究其实质,还是难以跳出他在引论中所批评的那种窠臼:"摹仿从前西文 Grammar 的'词类本位'的文法组织,……仅就九品词类,分别汇集一些法式和例证,弄成九个各不相关的单位……。"于是,该书除了六大句子成分和相应的图解法(Diagram)之外,真正关于句法的东西并不多。可见,少讲一点儿词法、多讲一点儿句法,在当时是不能也,非不为也。就是赵元任 1968 年出版的英文版《中国话的文法》,虽然采用结构主义的描写方法,但是仍然以词法为主、句法为辅。全书共八章,第一章序论,讲语

---

① 详见 Bloomfield (1933),Wells (1947),Hockett (1958)等;中文介绍和汉语层次分析方法问题,详见范继淹(1964/1983)。

法、语音、口语等问题；第二章讲句子、第五章讲句法类型，属于句法问题；第三章讲词和语素、第四章讲形态类型、第六章讲复合词、第七、八章讲各别词类，属于词法（形态）问题。难怪汤延池（1983）要批评："在语料的选择方面稍嫌守旧，而且在学术观点上依旧偏向以词为本位的传统语法分析……"（第 150—151 页）。倒是先前的吕叔湘（1942）《中国文法要略》（简称《要略》）和王力（1943）《中国现代语法》（简称《语法》），差不多在同时不约而同地尝试加大句法部分的比重。《要略》分上下两卷，上卷"词句论"占全书不到三分之一的篇幅，第一章讲字和词等词法问题，第二章讲词类及其配合关系，第三～五章讲叙事句、表态句、判断句、有无句；第六章讲句子和词组的转换，开了研究汉语句法结构变换关系的先河；[①]第七章讲繁句（包含多个词组的句子），第八章讲句法的变化；下卷表达论占全书三分之二以上的篇幅，把数量词、代词、方位词、时间词、否定词、助动词、语气词、连词及其用法，分别纳入数量、指称（包括有定和无定）、时间、正反·虚实、传信、传疑、行动·感情等意义范畴和离合·向背、异同·高下、同时·先后、释因·纪效、假设·推论、擒纵·衬托等意义关系中加以讨论。一般的语法书大多是从听和读等理解的角度来组织的，通常都以语法形式（结构、语序、虚词）为纲来说明其所表达的语法意义。由于汉语的结构和语序在当时的研究水平之下可说的实在不多，于是虚词等词法的内容就势必膨胀。像《要略》表达论中的这些内容，一般都是放在词法中讨论的。但是，吕先生在 F. Brunot（1922）*La Pensee et la Langue* 的影响下，别出心裁地从说和写的人的角度出发，以语法意义（各种范畴和关系）为纲来说明其赖以表达的语法形式。[②] 因此，虽然作者在写作上沿用的是前人写书讲虚词和句读的传统精神：类集用例、随宜诠释、稍加贯通，但是不仅给人耳目一新的感觉，而且确实对于读者理解和运用各种词类和语法格式有很大的帮助。[③] 至于王力先生，他在写作《语法》的时候已经认

---

① 参考朱德熙先生为商务印书馆《汉语语法丛书》所写的序，第 3 页。

② 详见吕叔湘先生 1982 年为该书写的"重印题记"，第 5 页。

③ 详见吕叔湘先生 1956 年为该书写的"修订本序"，第12 页。

识到：他在清华（国学）研究院做的论文《中国古文法》，除了死文法和活文法的分别、词有本性准性变性等说法颇有可取之处，其余就殊无可观。因为“当时的毛病是只知有词不知有句；只知斤斤于词类的区分，不知中国语法真正特征之所在；只知从英语语法里头找中国语法的根据，不知从世界各族语言里头找语法的真诠”（自序第2页）。于是，在他的《语法》中，第一、二章讲造句法，词类、词品问题也纳入其中，特别是提出了能愿式、使成式、处置式、被动式、递系式、紧缩式等句法格式；第三章讲语法成分，把系词、否定词、副词、记号（结构助词“的、所”、前缀“第”、后缀“子、儿、头”、时态助词“了、着”等）、语气词、连接词（结构助词“的”、连词“和”、介词“于、以”等）纳入其中；第四章讲替代法和称数法，把代词、数词等纳入其中；第五章讲倒装、省略等特殊形式，除了讨论复说、倒装、省略等句法问题外，把重叠、复合等构词法、乃至拟声、摹状等造词法的内容也纳入其中；第六章讲欧化的语法。到了丁声树等（1961）《现代汉语语法讲话》（简称《讲话》），句法的内容已经占语法的主导地位。《讲话》分别列专章讲句法结构、句子类型、主语、宾语、修饰语、补语、连动式、兼语式、连锁式、复合成分、复合句、问句等句法问题，就是讲词类也是着眼于词的用法（即句法分布）、讲语气词也是以疑问、祈使、测度、陈述、停顿等语用功能为纲来组织的。朱德熙（1982）共十八章，第一章讲语法单位；第二章讲词的构造，属于词法（形态）问题；第七～十二章讲主谓、述宾、述补、偏正、联合、连谓等句法结构，第十五章讲疑问句和祈使句，第十七章讲复句，第十八章讲省略和倒装，都属于句法问题。第三～六、十三、十四、十六章分别讲各种词类，但是作者强调“汉语不像印欧语那样有丰富的形态。因此给汉语的词分类不能根据形态，只能根据词的语法功能。……一个词的语法功能指的是这个词在句法结构里所能占据的语法位置”（第37页），所以完全是以句法结构（词组）为本位的。[①] 看来，从词法、形态本位向词组、句子本位转移是汉语语法研究的必然趋势。

---

① 关于在词组的基础上来描写句法、建立一种以词组为基点的语法体系的思想，详见朱德熙（1985）第74—79页。

顺便说一句，在大陆的现代汉语语法学界，前些年对于语法研究的本位意识明显地增强，提出了一些显然不同于朱德熙先生的"词组本位"的观点。比如，徐通锵先生提出了"字本位"的学说，邢福义先生提出了"小句中枢"的学说；此后，又有马庆株先生提出词和句子的"双本位"学说，以及大多数学者隐而不发的"无本位"思想。我们认为，只要不单纯是停留在喊口号、树旗帜的层面，从不同的角度思考语法研究的立足点，尝试建立新的组织语法知识的描写体系，肯定是有其积极意义的。在此，我们就不多作评论了。

3 语法分析的三个平面及其互动关系

上个世纪八十年代初，中国内地语法学界引进了语法分析的三个平面的观念。于是，原来许多纠葛不清的问题终于可以有一个名正言顺的说法了。比如，最著名的主宾语问题，到底是根据位置关系（在动词前或后），还是根据意义关系（施事、受事等），还是综合考虑两者，在五十年代是争论不清的。[1] 现在，朱德熙（1982）可以用三个平面的观念举重若轻地说：

> 主语和谓语的关系可以从结构、语义和表达三个不同的方面来观察。从结构上看，在正常的情况下，主语一定在谓语之前，两者之间的关系，跟其他各种句法结构比较起来，要算是最松的。这主要表现在以下两点上：第一，主语和谓语之间往往可以有停顿，而且后头可以加上"啊、呢、吧、嘿"等语气词跟谓语隔开。……第二，只要不引起误解，主语往往可以略去不说。……
>
> 从语义上看，主语和谓语的关系是很复杂的。拿动词组成的谓语来说，主语所指的事物跟动词所表示的动作之间的关系是各种各样的。有的主语指的是动作的发出者，即所谓施事；有的是受动作影响的事物，即所谓受事；有的是施事、受事以外的

① 详见1955年7月至1956年3月在《语文学习》上进行的关于主语宾语的讨论，文章收入《中国语文》杂志社编《汉语的主语宾语问题》，中华书局，1956年。

另一方，可以称为“与事”；有的是动作凭借的工具；有的主语表示动作发生的时间或处所。……注意不要把主语跟动作的施事混为一谈。

从表达上说，说话的人有选择主语的自由。同样的意思，可以选择施事作主语，也可以选择受事或与事作主语。……说话人选来作主语的是他最感兴趣的话题，谓语则是对于选定了的话题的陈述。通常说主语是话题，就是从表达的角度说的，至于说主语是施事、受事或与事，那是从语义的角度说的，二者也不能混同。（第95—96页）

朱德熙（1985：37）则总结性地指出：

进行语法分析，一定要分清结构、语义和表达三个不同的平面。结构平面研究句子里各部分之间形式上的关系。语义平面研究这些部分意义上的联系。表达平面研究同一种语义关系的各种表达形式之间的区别。这三个方面既有联系，又有区别，不能混为一谈。在上面提到的那些概念里，主语、宾语属于结构平面，施事、受事属于语义平面，话题、陈述属于表达平面。

上面朱先生所说的结构平面，一般称为句法平面（syntactic plane），表达平面一般称为语用平面（pragmatic plane）。按照我们的理解，语法理论模型中的语用平面的研究内容，可以更明确地界定为：研究语义相同或相近的各种句法格式在语用上的差别，包括它们对于不同的语境（context）的适应情况以及在会话涵义等推导意义方面的差别。

当三个平面的概念深入人心的时候，人们自然地会拿这三个平面跟传统的语音、词汇、语法三分格局和词法（形态）、句法二分格局进行比较。于是，疑问也就随之而来：跟语法分析关系密切的语音和词汇、特别是词法（形态），怎么在三个平面的分析框架中就没有地位了？其实，众所周知，语法分析的三个平面的观念是从符号学（semiotic）和数理逻辑（mathematical logic）中借来的。在这两门主要研究符号表达式的构造和推导关系的学科中，把对符号表达式的研究分成三个平面：（i）句法学（syntax），研究符号与符号之间的结构

关系，特别是什么样的符号表达式才是合式的(well-formed)；(ii) 语义学(semantics)，研究符号与所指(referent)之间的关系，特别是符号表达式为真的世界模型(即真值条件，truth condition)；(iii) 语用学(pragmatics)，研究符号与语境(包括说话人和听话人)的关系，特别是符号表达式在不同语境中的各种推导性的意义。既然符号学和数理逻辑以形式语言(formal language)的符号为起点，自然不会考虑自然语言(natural language)中的语音、词法等问题。因此，三个平面只是语法分析中一种观察问题的角度(perspective)，而不是语法分析的全部。

事实上，我们不仅应该分清语法的三个不同的平面，而且应该观察这三个不同的平面之间的互动关系。比如，像陈平(1994)那样，研究施事、受事等语义成分跟主语、宾语等句法成分之间的配位关系(argument selection)。并且，还可以引入语言类型学的视野，比较不同语言在配位方式上的共性和变异，整理出语义成分和句法成分在配位关系上的标记模式(markedness model)。[①] 从而，在世界语言的普遍性和差异性的可能范围这种广阔的背景上，来重新认识汉语的结构特点。更进一步，引进语法化(grammaticalization)这种动态性的概念，[②]来审视语法、语义和语用这三个平面之间的互动关系。特别关注不同语言中的有关语义和语用因素是怎样用语法形式来进行组织和编码的。这样，就可以既从共时角度出发，考察一种语义、语用现象怎样被语法形式进行编码；又从历时的角度出发，考察语法形式的起源及其虚化途径。从而打破共时研究和历时研究之间的藩篱，把语言的共时研究和历时研究沟通起来；彻底肃清索绪尔划分语言研究的共时平面和历时平面所带来的消极影响，推动语言研究走向更为全面、综合和多层面互动的道路。

---

① 比如，Comrie (1981) *Linguistic Universal and Language Typology*，通过跨语言调查发现：选择语义上的施事和语用上的话题作主语是一种语言普遍现象。详见中译本《语言共性和语言类型》第 13 页，沈家煊译，华夏出版社，1989 年。另外，沈家煊(1999)对于汉语语法中的有关标记模式进行了探讨，值得参考。

② 详见 Heine，Claudi and Hunnemeyer (1991)，Hopper and Traugott (1993) 和 Bybee，Perkins and Pagliuca(1994)等。

4 语法和语音的互动关系的初步探索

在汉语研究中，学者们较早地认识到语音和语法有互相制约的作用。比如，林焘(1957)分别考察了现代汉语趋向补语、可能补语、程度补语和少数结果补语中轻音现象所反映的语法和语义问题，发现语音格式的不同对语法和语义有直接的影响。例如：①

(1) a. 想了很久，我才想·起·来了。(趋向补语——引申意义)

b. 天气不早，我想起·来了。(主要动词)

(2) a. 日子隔得太久，我想·不起来了。(可能补语——引申意义)

b. 今天我有点不舒服，我想不起·来了。(主要动词)

在上例中，词语序列"想(不)起来"是靠着不同的轻音现象来分别它们的语法作用和意义的(第4页)。再比如，"死、开、到、着"等由于语法作用和语音不同，每个词至少具有三种不同的意义。例如：

(3)

动词	非轻音补语	轻音补语
他死了	看死了	乐·死了
开门了	想开了	走·开了
到北京	想到了	提·到你
火着了	买着了	打·着了

关于这些具有对立价值的例子，林先生富有洞察力地指出：

> 动词和非轻音补语在意义上的不同决定于语法作用的不同，非轻音补语和轻音补语在意义上的不同决定于声音的不同。这种现象最足以说明语音和语法以及语义之间的密切关系，也正可以提醒我们绝对不能把语言的这三方面割裂开来孤立地进行研究。(第21—22页)

① 在汉字前面加"·"号表示轻音。

林焘(1962)指出,研究一种语言的句法结构,主要是从词和词之间的结构关系入手。这种结构关系有时能从语音现象中(包括语音的停顿、高低、轻重等)反映出来(第23页)。基于这样的认识,他对现代汉语轻音和句法结构的关系进行了更全面的考察。根据分布和功能,他把普通话的轻音分为两类:语调轻音和结构轻音。语调轻音跟相关上下文中的语调重音相对立,表示不同的语气,跟语言的结构层次没有直接的关系。例如:①

(4) a. 他·是学生。(一般叙述)

b. 他′是学生。(我的看法并不错)

c. 他″是学生!(你别以为他不是学生)

三句话结构完全相同,“是”的三种读法只是表达了三种不同的语气。结构轻音跟语言结构或意义关系密切,在同样的上下文中一般没有重音跟它对立。比如,上文的例(1)(2)。再如:

(5) a. 刚安静了一会儿,你们又说·开了。

b. 事情说开了,咱们俩心里也就痛快了。

其中的“说开”的语音和意义都是不同的。结构轻音具有后附性的特点,它可以帮助我们确定语言的结构层次。例如:

(6) a. 他·的|书 ～ b. *他|·的书 ～ c. *他|·的|书

(7) a. 住·在北京、生·在一九六二年、跑·到屋里、写·到晚上十二点

b. 放·下书包、跑·进屋里去、借·来一本书

c. 送·给你、借·给他、借·给一个人

林先生指出,在分析语法层次时,除非有特殊的理由,不应该把由结构轻音构成的语音层次任意打乱。……像“们、的、地、得、了、着、过”等语法成分永远轻读,正是划分层次的标志(第36页)。因此,(6a)这种分析最合理。考虑到轻音在语法结构中的作用,“动词+在/到/给”应该分析为一个直接成分,其结构关系跟“动词+下/进/来”一

① 在汉字前面加“′”号表示一般重音,加“″”号表示强调重音。

样，是述补结构（第 40 页）。至于“院·里、墙·上、年·下”等“名词+方位词”结构，考虑到其语音、意义和结构的各个方面，不应该分析成偏正结构，而是可以看成是名补结构（第 44—48 页）。这种对轻读成分的附着现象（clitics）及相关结构的研究，理论语言学界大概是在 Zwicky (1977) 之后才被重视和得到广泛的研究，并形成蔚为大观的形态句法学（morphosyntax）的。

更难能可贵的是，林焘（1962：27—29）在考察轻音在语音结构中的地位时发现：

> 上声在一般轻音之前只读成“半上”[21]，它后面的轻音音高是[4]，两个音节恰好共同构成一个全上声的调值[214]。去声在轻音之前只读成[53]（或[52]），它后面轻音音高是[1]，两个音节恰好共同构成一个全去声调值[51]。这种变化不只能说明轻音在语音结构中的地位，而且也可以看出汉语声调的调值有超出一个音节的范围而把后面轻音音节包括进去的趋势。
>
> 第一个上声音节有时可以仍根据后面轻音原来的声调来变调，有时又可以不管后面轻音音节，直接和轻音后面的其他音节发生变调关系。例如：
>
> (8) a. 你·们　　b. 你好　　c. 你·们好
>
> 买·了　　买米　　买·了米
>
> a 栏的第一个音节在轻音字前，所以读半上；b 栏的第一个音节在上声字前，所以读阳平；
>
> c 栏的轻音之后跟着一个上声字，第一音节可以读半上，也可以读阳平。……普通话三音节连续快读时，如果第一音节是阴平或阳平，第二音节是阳平，则第二音节可以变调读阴平。……第三音节是轻音时……第二音节都不变调……如果轻音音节之后紧跟着另一个有声调音节同时快读时，则第二音节仍然可以变调。例如：
>
> (9) a. 非常·的好 /fei^{55} chang^{35} de hao^{214} /→
>
> / fei^{55} chang^{55} de hao^{214} /
>
> b. 说服·了人/shuo^{55} fu^{35} le ren^{35} /→

/ shuo^{55} fu^{55} le ren^{35} /

而且轻音的音高也上升接近[5]。这说明我们在说这句话时，已经倾向把这个轻音音节和它前面的音节看成共有一个调值了。

众所周知，从 Goldsmith (1976)提出自主音段音系学(autosegment phonology)理论以后，许多语言学家很坚决地相信：声调是一种独立于它所搭乘的音段的自主音段(autosegment)。① 于是，像吴语等汉语南部方言中连读变调(tone sandhi)时强读音节的声调保持并扩散到整个变调域(sandhi domain)的现象，就可以用弱读音节的声调删除和强读音节的声调扩散和连接到其他音节上等观念来描写和解释。没想到林焘(1962)在讨论以北方官话为底子的普通话时，已经涉及这种现象，并正确地概括为："汉语声调的调值有超出一个音节的范围而把后面轻音音节包括进去的趋势"(第 27 页)和"这说明我们在说这句话时，已经倾向把这个轻音音节和它前面的音节看成共有一个调值了"(第 29 页)。遗憾的是没有人能以此为观察和思考的起点，演绎出类似形态句法学、自主音段音系学那种具有解释性和普遍性的语言学理论。说起来大有令人扼腕之感，个中原因谅非三言两语所能道明，在此姑且按下不表。

吕叔湘(1963)考察了现代汉语单双音节词的语法功能的差异问题，发现：在现代汉语中，(i) 单音成分的活动是常常受到一定的限制的。因此，经常通过附加没有多少意义(失去原有意义，没有对立、区别作用)的"老、小、子、儿、头"、方位词、或同义并列、重叠等构词手段，来造成双音词。例如：

(10) 老虎～小老虎　小偷～*大偷　石头　心里　衣服　灯火　星星

(ii) 在三音节和四音节的语音段落里，有单音节和双音节的搭配问题。例如：

(11) 进行调查～*进行查　管理图书～*管理书

① 详见 Bao (1999), p. 5—7; Chen (2000), p. 57—63.

钢铁生产～＊钢生产

伟大人物～＊伟大人～伟大的人

强大的国家～＊强大的国

其中,吕先生特别敏锐地发现:三音节的语音段落,偏正式合成词,2＋1式(动物学、示意图)比1＋2式(副作用、手风琴)要多得多;动宾组合,1＋2式(买东西、写文章)多于2＋1式(吓唬人、糟蹋钱)。至于其中的原因,吕先生说:前者跟在前或在后的单字的性质和可以这样用的单字的数量有关系,后者跟常用动词中单音的较多有关系;不过,他都极为谨慎地指出:是否完全由于这个因素,还需要进一步分析。二十多年后,吴为善(1986、1987、1989)和陆丙甫(1989)、张国宪(1989)等继其余绪,进一步探讨汉语音节组合的规律及其背后的原因。其中,吴为善(1986)指出:最常用的动词大多为单音节(约70%),名词大多为双音节(约85%),使得由它们构成的动宾结构大多为1＋2的模式。他还很好地得出了较具概括性的结论:词语搭配的选择性,除了语法上和语义上的制约,还有语音方式上的限制。但是,他没有说明为什么三音节的名词性偏正结构大多为2＋1模式。这个问题在吴为善(1989)中也许可以找到一定的解释:三音节段的动宾结构在意义上都是两个概念的组合(运粮食、缝衣服),它们之间的关系较松;而偏正结构实际上相当于一个复合词(象牙筷、防风镜),它们大多是一些事物的名称,表示的只是单个的概念,内部比较紧密。其他学者的发音和听辨实验也证明:在1＋2中,前面的单音节跟后面的双音节结合较松;在2＋1中,后面的单音节跟前面的双音节结合很紧。他由此得出推论:后置单音节具有粘附性,前置单音节具有相对独立性。并作出更为大胆的设想:一定的语义、语法组合总是选择适当的语音组合形式,使两者一致起来,就像在三音节段里,动宾往往选择1＋2,而偏正往往选择2＋1。陆丙甫(1989)尝试用结构(汉语是核心在后结构占优势)、节奏(汉语三音节是1＋2优于2＋1)、松紧(核心在后结构紧于核心在前结构)、轻重(有无轻读音节)等概念,来解释为什么汉语中某些格式是不合格的。张国宪(1989)则专门考察“动＋名”结构中单双音节动作动词功能的差异,特别是构成三音节和四音节组合时,其结构关系(动宾还是偏正)、动

词和名词之间在音节、语义等方面的选择限制。应该说是有了一个良好的开端,可惜的是这种研究大都比较零碎、缺少系统性,也没有联系现代音系学和现代各种相关的句法理论;因此,热闹了一阵子之后,便难以为继,只得沉寂下去。

5 语法和语音的互动关系的系统研究

对汉语中语法和语音之间互动关系的较为系统一点的研究,是在生成语法和生成音系学的理论背景上展开的。并且,呈现出明显的不对称性,表现为:关于语法对语音的制约作用的研究较为充分,而关于语音对语法的制约作用的研究尚嫌薄弱。具体地说,由于汉语是一种有声调的语言(tonal language),并且汉语各方言都有声调在特定的上下文语境中的读音变化(tonal alternation)现象,即连读变调(tone sandhi)。严格地说,连读变调是一种发生在词或语素的交接处的"语素—声调音位"变化(morphotonemic alternation);它有时不能单纯从语音条件上作出描写,而是要联系到相关的形态条件或句法条件、甚至是语义和语用条件。其中,最著名的是连读变调的范围(即变调域,sandhi domain or scope)问题。众所周知,跟绝大多数的局限于一处的音段现象不同,声调的作用范围是长距离的,有时甚至横跨整个短语或句子。于是,要想准确地界定变调域,就必然会引起诸如音系和语法结构的关系等有趣的问题。① 比如,在浙江汤溪话中,短语和合成词的重音(stress)位置不同,从而造成了不同的连读变调的结果。例如:从本调来看,"炒"是降调(HL),"饭"是升调(LH)。当这两个语素组成动宾短语时,重音在宾语"饭"上,"炒"的声调删除;当这两个语素组成偏正合成词时,重音在修饰语"炒"上,中心语"饭"的声调删除,并且"炒"的声调扩散(spread)到"饭"上。上海话中"炒饭"的情况与此相似,作偏正式合成词时,"饭"的声调(LH)删除,"炒"的声调(MH)系联(association)到"饭"上;作动宾短语时,可以保持本调不变,或者"炒"的声调简化(simplification)为

① 参考 Chen (2000) 的 Preface, p. xi, xiii。

H。值得注意的是,汤溪话中的变调规则主要是直接跟形态—句法结构(morphosyntactic structure)相联系的,表现为:在一个短语中,只有最右边的词的起首音节保留其原有的词汇声调并向右扩散到整个组合。上海话的情况比汤溪话复杂,是以重读音步(stress-foot)这种节律(metrical)单位为连读变调域的。尽管如此,上海话的节律组织还是跟形态—句法直接相关的。比如,端木三的一系列研究显示上海话的节律规则在语素、词/合成词、短语三个平面上是不同的。[①]这说明像连读变调等语音现象不仅不是无视句法的(syntax-blind),而且有时是句法敏感的(syntax-sensitive)。于是,在描写语音现象、总结变调域等音系规则时,就不可避免地要用到许多句法学的概念,甚至是一些极为抽象的句法概念。比如,在描写丹阳话的变调域时,端木三用基于重音的节律分析法(stress-based metrical analysis);而张洪明则沿着基于句法的路子(syntax-based approach),用成分统制(c-command)这种抽象的概念来定义变调域:[②]

α向右扩散到β,当且仅当β被α成分统制。

众所周知,当代生成语法的有些句法概念是极其抽象的;有时,用这种过于抽象的句法概念来描写音系规则,会掩盖音系限制条件背后的动因(motivation)。比如,南通话的韵律组织(rhythmic organization)基本是依赖于结构的(structure-dependent)。因此,同样是四个音节,"(红十字)(会)"是两个音步、两个变调域,"(波里)(维)(亚)"是三个音步、三个变调域。之所以不同,都有结构上的原因。但是又不能简单地说直接成分(IC)必须属于同一音步。因为,存在着"(红十)(字)"(两个音步、两个变调域)这种不顾直接成分结构的韵律组织。为了解决这个矛盾,敖小平提出了下列十分笨拙的规定:

---

① 详见 Chen (2000), p.297—299, p.306—316, p.88。

② 详见 Chen (2000), p.331—335。更全面地说,张洪明(1992: 224)指出,变调域是由成分统制和论旨关系(thematic relation)决定的。有关的评论请看 Duanmu (1995), p.228—229。

语素完整性限制(Morpheme Integrity Constraint):

一个管辖(dominate)α和β的音步,管辖所有成分统制β且不成分统制α的γ,如果γ在一个可以成为音步的δ之前。

这里的α,β,γ,δ指一个线性的音节序列,比如"红十字会"。显然,这种限制用了复杂的关于成分统制关系的规定,其实是专门为了能够说明"红十字会"一类韵律组织,同时又绕过"红十字"这种违背直接成分结构的韵律组织。至于这种限制背后的动因是不清楚的。相反,陈渊泉为此提出了具有独立动因的原则:

不要骑跨(No Straddling):

直接成分必须是一个音步之中的伙伴(IC must be foot-mates)。

至于"红十字"的韵律组织违背这一限制依然成立的原因是,其他可能的韵律组织将违背更多或更重要的限制;根据优选论(optimality theory),采纳目前这种韵律组织虽然是美中不足,但相比之下仍是最优的。[①]

特别有意思的是,韵律组织有时会直接受到语义、语用因素的影响。比如,张正生和石基琳曾经分别指出,说话人往往会通过在一个成分之前放置一个强调边界(emphatic boundary,记作!),来标志这个成分处于焦点或对比地位。这种边界就像是语调短语(intonation phrase,记作 IP)边界,成为音步组织的一个新的参照点。于是,通常的韵律组织可以给出一个无标记的解读(reading),有由语用决定的强调和对比标记的韵律组织给出一个有标记的解读。例如:

(1) 只〔买 股票〕,不〔卖 股票〕。

a. (s 3) (3 4) …

b. (3)! (s 3 4) …

其中,s 代表从第三声通过连读变调而派生出来的第二声,3 表示第三声,4 表示第四声。(1a)是无标记的解读,(1b)是有标记的解读。

① 详见 Chen (2000), p. 356—360。

韵律结构和语法结构之间的互动关系，在这里表现得可谓淋漓尽致。Chen (2000: 404)为北京官话(Beijing Mandarin)的变调域所受到的制约因素排列成如下这种层级序列：

{不要骑跨，语调短语界限}>{二元性}>
{有限性}>{一致性}>{从左向右}

因为，话语首先划分成语调短语，然后音节在语调短语的范围内组合成最小的韵律单位(minimal rhythmic units，简称 MRUs)。这种 MRUs 成为强制性的连读变调域。所以，语调短语界限和保持直接成分在一个音步之内的不要骑跨是最优先的限制。二元性要求一个 MRUs 最少有两个音节，有限性则要求一个 MRUs 最多只能有两个音节。一致性要求组成一个 MRUs 的成员是形态—句法上关系紧密的伙伴，以保证韵律跟句法的和谐。从左向右说的是组织 MRUs 的方向和顺序。[①] 由于语调短语的界限有时要受到语用的影响(如 1b 所示)，因而上述六个限制条件中，竟然有三项是跟语法(包括语义和语用)相关的。语法对语音的制约关系，在这里得到系统而清晰的反映。

至于语音对语法的制约关系，就没有这么直接和系统了。根据 Bloomfield (1933: 163—165)，一个语言中形式的有意义的配置方式(meaningful arrangements)构成了这语言的语法。语言形式的配置有四种方式：(a)次序(order)，比如，John hit Bill 不同于 Bill hit John，Bill John hit 则不是英语的句子形式；(b)节律(modulation)，比如，"John." "John!" "John?"通过语调音高变化来表示陈述、回答、疑问等意义差别；(c)语音改变(phonetic modification)，比如，run ～ ran，keep ～ kept；(d)选择(selection)，比如，drink milk ～ watch John，fresh milk ～ poor John，John runs fast ～ the boys run fast，这种搭配不能任意替换。Chao (1968) 指出：这四种方式在汉语语法里的作用有大有小。在近代汉语里，节律和语音改变的作用不太重要，次序和选择在语法安排上起主要的作用。……节律指轻

① 详见 Chen (2000)，p. 371—372，380，404。

重、停 顿、语调等方面的差别(吕译本第 9 页)。的确,在传统的结构主义语法框架中,像直接成分之间的结构关系、各直接成分的音节和整个组合之间的合格性之间的制约关系是不太容易处理的。等到生成音系学诞生,轻重、音步、韵律的冲突和消解(metrical clash and resolution)等概念和相应的分析方法出现,比较系统的关于语音对语法的制约关系的研究终于有了技术上的可能性。比如,冯胜利(1997,2000)等著作,用韵律音系学的理论和方法来考察和分析汉语语音对语法的制约关系,探索句法与韵律的相互作用的规律,并尝试建立汉语韵律句法学。讨论到的问题包括:汉语的自然音步是什么?是怎样构成的?有没有必要在单音节词和句子之间建立韵律词(prosodic word)这一单位?不同类型的句法结构(比如,述宾结构和偏正结构)在单、双音节的搭配上有什么限制条件?为什么?能否和怎样用韵律要求来解释“把”字句、“被”字句、主题句等句子中宾语位置的移动,动词之后的介宾结构中的介词贴附在动词上,历史上介宾结构位置从动词后向动词前的转移、SOV 结构向 SVO 结构的转移、以及“被”字句和“把”字句的产生和发展等历史句法问题?其中,比较引人入胜的是希望用音步、轻重及其造成的韵律格式等概念来解释下列不平行的现象:

(2) 种植花草~*种植花/草~种花草~种花/草

(3) 阅读报纸~*阅读报~读报纸~读报

(4) 喜欢钱财~喜欢钱~爱钱财~爱钱

(5) 皮鞋工厂~皮鞋厂~*鞋工厂~鞋厂

(6) 煤炭商店~煤炭店~*煤商店~煤店

如果假设在韵律上双音节比单音节重或突出,或者说较重的成分采用较长的词形;并且,相信重音由句法关系决定,那么就可以引入一条辅重原则(Non-Head Stress principle, NHS):在韵律上,辅助成分(即非核心成分,包括论元和附加成分)应该比核心成分更重、更突

出(prominent);于是,上述现象就比较容易解释:[1]述宾结构的核心是述语,辅助成分是宾语;因此,述宾结构在韵律上应该是抑扬格(iambus)。于是,1+2 式的述宾结构符合抑扬格这种韵律限制,而 2+1 式是扬抑格(trochaics),不符合述宾结构的韵律要求。至于"喜欢"的"欢"是个轻声字,整个双音节词是一个打了折扣的残音步(defective foot),其长度跟单音节差不多;也就是说,"喜欢钱"这类形式不是抑扬格,所以并不违反述宾结构的韵律限制。[2] 这是对(2)—(4)的解释。至于(5)(6)是偏正结构,根据辅重原则,应该是扬抑格,所以 1+2 式不符合这种韵律要求。

上述这种韵律句法学解释看上去简单明了,又很能说明问题。但是,如果像 Lu and Duanmu (1991)那样,坚持认为:对词的长度的选择取决于重音,带有重音的词不能比不带重音的词短;那么,就不足以解释吕叔湘(1963)中碰到的一些问题:(1) 偏正结构固然以 2+1 式为主流,但是 1+2 式也为数不少;述宾结构固然以 1+2 式为主流,但是 2+1 式也为数不少。例如:

(7) 副作用 手风琴 大面积 小规模 新衣服 老工人 长短裤

(8) 浪费钱 需要纸 爱护人 采购米 产生电 答应去 分裂党

于是,前述的韵律限制就不像是强制性的(obligatory)了。但也不像是任选的(optional),因为有"鞋工厂、种植花"等不合格的形式存在。

---

① 冯胜利(2000)从"句子重音在句末的最后一个短语中实现"这种普遍语法原则出发,来推导出"宾语永远比动词强"(第 115 页)。因为汉语没有形态,动词没有定式(finite)和不定式(infinite)之别;并且,述宾结构的韵律限制在述宾结构嵌套进其他词组中时依然有效(如:种植花草的季节～*种植花的季节)。所以,我们完全可以在词组平面上来研究词语组合的韵律限制。关于句法决定重音和辅重原则,详见 Duanmu and Lu (1990), Lu and Duanmu (1991), Duanmu (1990, 2000)。Duanmu (1995: 255)特别指出,根据 Cinque (1993),他假定:短语的重音是循环指派的,宾语的重音比动词强,谓语的重音比主语强。Chen (2000: 257, 491)也有精彩的讨论,可以参看。Chen (2000: 500)特别指出,Cinque (1993)的普遍重音规则给出的优先顺序是:补足语(Comp) > 核心(head) > 指示语(Spec),Duanmu (1990)的辅重原则并不区分论元和附加成分(adjunct);而他则通过汉语方言的例证给出了这样的可重读层级(stressability hierarchy):附加语(adjunct) > 论元(argument) > 核心(head)。

② 关于对"残音步"的分析,详见冯胜利(2000),第 119—120 页。

对于“大苹果、布手套”一类 2＋1 式偏正结构的合格性，Duanmu (2000)提出了一种解释：因为像这里的“大、布”等在长度上不能变通，即它们没有同义的双音节形式。我们认为这种解释说服力不强。比如，“大型苹果”固然是语义怪异，但是“大号苹果、大的苹果、大大的苹果、棉布手套、布制手套”等符合韵律要求的同义形式是现成的。我们建议，可以像 Chen (2000：222，257，500)那样，在辅重原则 NHS 之外，引入 Prince (1990)提出的分量适合重音原则(weight-to-stress principle，WSP)，即重音落在韵律分量重的成分上比落在韵律分量轻的成分上更和谐(harmonic)。这样，分量适合重音只是一条优选论原则，不具有强制性；更何况实现韵律分量重，除了音节数量之外，还有拉长元音等手段可资利用。(2) 符合句法、语义选择关系，并且明明符合韵律限制的格式反而是不合格的。例如：

(9) 互相埋怨～＊互相怨～互相咬
共同使用～＊共同用～共同做
日益增多～＊日益多～?日益少

(10) 伟大人物～＊伟大人～伟大的人
强大的国家～＊强大的国

(11) 中药西药～中西药
大事小事～＊大小事～大小事务

(12) 编辑和出版刊物～＊编和出刊物～编刊物和出刊物

要解释这些现象，就需要我们考虑更多的因素及其复杂的相互作用机制。

最后，韵律句法学上建立起来的音步的概念，应该跟音系学上关于北京话连读变调域的音步或最小韵律单位(MRUs)进行比较。比如，冯胜利(2000)提出：一个在韵律节奏中可以独立的基本单位是韵律词，韵律词是一个最小的语流片段。他相信 McCarthy & Prince 的说法：人类语言中“最小的能自由运用的韵律单位”是“音步”；因此，主张用音步来确定韵律词。他指出，音步由音节组成，韵

律词则由音步来实现。为此，他提出了下列汉语音步组成的规则：[①]

(i) 汉语自然音步的音节“小不低于二，大不过三”；因此，

(ii) 单音节形式不足以构成独立的音步，如：法(国)、美(国)；

(iii) 两个音节组成一个独立的音步，如：巴西、古巴；

(iv) 三个音节也组成一个独立的音步，如：加拿大、墨西哥；

(v) 四个音节必须分成[2＃2]格式，如：斯里/兰卡、坦桑/尼亚；

(vi) 五个音节必须分成[2＃3]格式，如：阿尔/巴尼亚、加利/福尼亚；…

冯胜利(1997)参照陈渊泉、石基琳的办法，建立了划分句子音步的程序：

(i) 先按直接成分分析法切分句子，如：校长||想请小王|吃晚饭；

(ii) 再从右向左系联各成分中的双音步，如：(校长)||想请(小王)|吃(晚饭)；

(iii) 剩余的单音成分仍系联成双音步，如：(校长)||(想请)(小王)|吃(晚饭)；

(iv) 不成双音步的单音成分系联到邻近的音步上，根据其句法关系决定左附或右附；如：(校长)||(想请)(小王)|(吃(晚饭))。(第23页，注4)

这里的音步，说得朴素一点儿就是：最小的两头可以有明显的停顿、或短暂的间歇的音节群。根据我们的体会，北京话的连读变调域也是最小的两头可以有明显的停顿、或短暂的间歇的音节群。正因为处于一个音步中的音节群中的各音节之间没有间歇，所以变调成了一种达到协同发音的调节机制。这样说来，音步或韵律词的概念跟

① 详见冯胜利(2000)，第77—80、93—98页；部分叙述有改动，例子有所增益。

最小韵律单位(MRUs)的概念应该是吻合和协调的。[①] 比如,Chen (2000: 367)提出的北京话的最小韵律单位(MRUs)的组织原则是:

(i) 二元性,一个MRUs至少是双音节的;
(ii) 有限性,一个MRUs至多是双音节的;
(iii) 从左向右,MRUs是从左向右组织起来的。

这跟冯胜利(2000)的音步组成的规则大致相似,只是音步组织的方向正好相反。[②] 但是,对一些多音节组合,根据冯胜利(1997,2000)得出的音步划分,是明显地不同于其连读变调域(即MRUs)的。例如:[③]

(15) 纸老虎跑	(16) 找胆小鬼	(17) 我往北走
a. (3 s s 3)	(3 s s 3)	(s 3) (s 3)
b. ( )( )	( )( )	( )( )

由于汉语允许单音节构成一个独立的退化音步(degenerate foot),因而从理论上讲b种音步组织是合理的。但是,a这种连读变调的语音事实却自成一格,并且有理可循。比如,(15)—(17)的b种读法将同时违背二元性和有限性两项限制,(16)b还违背了从左向右这项限制;而(15)(16)的a种读法只违背有限性一项限制;(17)的a种读法只违背不要骑跨一项限制,但介词"往"作为一种附着成分

① Chen (2000: 505)在全书结语(concluding remarks)中说:官话中的最小韵律单位(MRU)比较特别,它既可以是比词汇小的片段(sublexical fragment),也可以是整个词、一个短语、多个小句的结构(multiclausal construction),或者甚至完全是一个非结构成分(non-constituent)。因此,它不适合如下常规的韵律层级:{音步(foot),音系词(phonological-word),附着群(clitic group),音系短语(phonological-phrase),语调短语(intonation phrase)};但是,它作为一种特殊的韵律单位而独立存在。如果有什么区别的话,作为一种诗歌韵律分析的单位,MRU可以在音步中发现它自己的最接近的相似物。

② 心理语言学的证据显示,言语组织(即音系编码)有从左向右的倾向;并且,在把音节组织成音步的时候,也是从左向右占优势。参见Chen (2000: 119)及其所引的文献。另外,Shih (1986)也主张:从左向右把挂单的音节(unpaired syllable)系联成二元音步(即双音节音步),除非它们是朝相反的方向分枝的。转引自Chen (2000: 374)。

③ 关于这些组合的连读变调域及其限制条件,详见Chen (2000) Chpt. 9: Minimal rhythmic unit as obligatory sandhi domain, pp. 369—370, 400—401。变调域显然是跟音步相关的,比如Shih (1986)主张,官话第三声的变调域是音步,参见Duanmu (1995: 257)。

(clitics),它必须向左贴附(cliticize leftwards)到前面的音节上一起构成一个语音词(phonological word),这种强制性的音系学限制足以抵销不要骑跨限制。面对这种句法和语音错配(syntax-phonology mismatch)现象,我们必须思考:韵律句法学上的音步概念应该向句法倾斜还是向语音倾斜?说到底,应该怎样既有助于说明句法问题,又在一定程度上照顾到语音现实?

6 词库结构和句法操作的互动关系

在传统的语言研究中,人们相信:语法是语言的结构规律,词汇是语言的建筑材料。因此,在语法学中并没有词汇的地位。到了生成语法时代,语言学家以探究人类语言知识的组织方式并为之建立理论模型为己任,语法研究的对象扩大为整个的人类大脑中的语言知识。于是,语法除了包括其核心内容句法之外,还包括词汇、语义和语音。比如,Chomsky (1965)认为语法有句法、语义、音系三个组成部分。其中,句法部分包括短语结构规则和转换规则,短语结构规则和词库(lexicon)构成句法的基础部分。把取自词库的词项插入表示短语结构的树形图就得到句子的深层结构,再输出到语义部分就得到句子的语义表达式。对深层结构进行转换操作就得到句子的表层结构,再输出到音系部分就得到音系表达式。这就是所谓的标准理论。到了修正的扩充式标准理论,改为表层结构同时向音系部分和语义部分输出,词库和短语结构规则的关系则没有改变。既然词项是短语结构规则的输入,那么词项之间的搭配和组合信息越详细、越具有结构性就越有利于说明句法组合的合格性问题。比如,动词能不能带宾语,能带几个宾语。进一步的要求就是:(i) 动词能跟几个名词性成分发生句法、语义联系,这就是配价语法研究的主题;(ii) 这些名词性成分相对于动词分别充当了什么语义角色、在表层结构中分别能充当什么语法角色,这就是格语法的核心内容。Chomsky (1982)所建立的GB理论,已经很好地把这种知识作为记载在词库中的动词词条之下的词汇、语义特征;还假定:一个动词所必备的论元构成了动词的论元结构(argument structure),基础句式

是动词的论元结构的一个投影。在此基础上，建立起投射原则、论旨准则等 GB 理论的原则系统。至于 Jackendoff (1990)、Grimshaw (1990)等，则有意提高论元结构在语法理论模型中的地位，尝试把论元结构当作是介于词库和深层结构之间的一个独立的语言知识的表达层次，并着力探索论元结构的内部结构及其运作机制。更有甚者，Levin & Rappaport (1995)假设，在词库和句法表达(syntactic representation)层面之间有两个界面：(i) 词汇语义表达式(lexical-semantic representation)，(ii) 词汇句法表达式(lexical-syntactic representation)，也称为论元结构(argument structure)；词汇从词库到句法层面要先经过词汇语义表达式，再经过词汇句法表达式。某些词汇经过这两个层面可以衍变为新的词汇，如非宾格动词(unaccusative verb)、中间动词(middle verb)等。这是理论语言学界对词库和句法结构的互动关系研究的一个重要的方面，下面检讨一下汉语语言学中相应的研究和理论分歧。

汉语语言学界从上个世纪八十年代以来，先后对动词、形容词和名词的配价情况作了比较仔细的研究；又对述补结构、述趋结构等动词性结构的配价情况进行了初步的研究。并且，自觉地在论元结构理论的框架下对配价研究进行改造。后来，沈家煊、张伯江等先生又在 Goldberg (1995)的句式语法(Construction Grammar)思想的影响下，著文强调句式也有指派论元的功能、句式配价比动词配价更重要，希望以此来解释跟动词的论元结构不相符合的句式构造。例如：

(1) 王冕七岁上死了父亲。

(2) 我吃了小明一个苹果。

在上例中，一元动词“死”带了两个论元，二元动词“吃”带了三个论元。如果认为动词的论元结构决定了以它为谓语核心的句子的构造，那么就必须假设这里的“死、吃”发生了变价。这个口子一开，其结果会导致词无定价，从而宣告词汇主义(lexicalism)的破产。如果认为句式是一种独立的语法实体，可以容纳有关动词进入其中；那么，上述例子就比较容易解释。但是，随之而来的问题是：怎样说明不同的句式对有关动词的选择限制条件。比如，双宾语句为什么既

能容纳"给"类三元动词,又能容纳"吃"类二元动词;到底具有什么样的句法、语义特征的动词可以进入双宾语句?真理也许就在两极之间,采取动词的论元结构跟句式的论元结构互动的观念,也许可以较好地解决上述问题。[①]

下面,我们通过几种对于同一个具体现象的不同的处理方案,来看词库结构与句法操作的互动关系。众所周知,现代汉语中的工具、处所、方式、目的、共事等论元角色一般只能通过介词的引导而作状语;但是,有时它们中的一部分却可以有条件地直接作宾语。例如:

(3) 用毛笔写字 → 写毛笔　　用烟斗抽烟 → 抽烟斗
　　在食堂里吃饭 → 吃食堂　　在地板上睡觉 → 睡地板
　　用花腔唱歌 → 唱花腔　　用魏碑体写字 → 写魏碑体
　　为买带鱼排队 → 排带鱼　　为买钢材奔跑 → 跑钢材
　　跟日本队打球 → 打日本队　　跟野女人睡觉 → 睡野女人

这种现象,邢福义(1991)称之为"宾语代入",即其他成分代入常规宾语的位置。袁毓林(1998)从施事、受事等语义成分跟主语、宾语等句法成分配位的角度,假设存在着述题化这种调整配位的语法机制来作出解释。大意是:施事等主体性论元通常实现为句子的主语,受事等客体性论元通常实现为句子的宾语,工具、处所、方式等环境性论元通常实现为句子的状语,这是无标记的配位方式;但是,客体性论元和环境性论元可以通过话题化(topicalization)这种语法过程而实现为主语,主体性论元和环境性论元可以通过述题化(rhemization)这种语法过程而实现为宾语,这是有标记的配位方式。启动述题化这种语法过程的动因是说话人想让主体性论元和环境性论元成为句子的语义重心,所以强行占据宾语这个句子的常规焦点的位置。同时还假设,伴随着述题化这种语法过程,成为宾语的主体性论元和环境性论元在语义上经历了受事化的过程,即包含〔+受动〕这一动态的语义特征(第135—142页)。这种处理方案的缺陷是显而易见的。因为,述题化的语法过程,隐含了主体性论元和环境性论元向宾

① 关于这方面的讨论,详见袁毓林(2002a)及所引文献。

语位置移动这种句法操作；于是，在理论上就必须回答两个问题：(i) 向后移动的成分有没有留下语迹(trace)，如果有，那么后移成分怎么管辖其语迹，从而使句子获得正确的语义解释？(ii) 常规的受事宾语到哪儿去了，有没有留下语迹？如果有，那么由什么成分来管辖？从GB理论的眼光来看，向后移位这种句法操作是根本不可能的，它直接违背了投射原则和论旨准则。①

这些问题，冯胜利(2000)用焦点韵律迫使下的核心词移位这种理论假设，作出了很好的处理。其大意是："写毛笔"的底层结构是"用毛笔写字"，后者的构造是：②

(4) $_{VP}$[V(用)＋$_{VP}$[NP(毛笔)＋$_{VP}$[V(写)＋NP(字)]]]

当动词"用"和常规宾语"字"不出现的情况下，动词"写"通过从核心到核心的移位(head-to-head-movement)而进入"用"的位置(第171—179)。但是，这种分析还是留下了一堆问题：(i) 这个不出现的常规宾语是什么性质的空语类，要不要受到管辖，由谁来管辖？(ii) 引进工具论元的动词"用"是怎样"非音化"的？冯胜利(2000：176)说"表使动的动词在句法上可以'非音化'形式出现，因此表工具的动词也可以'非音化'形式出现"，这种推论的逻辑基础何在？(iii) 像(4)这种结构分析是特设的(ad hoc)，专为核心移位而假设的。(iv) 除了工具论元之外，还有处所、方式、目的、共事等环境论元，岂一个"用"字所能了得？(v) 根据"辅重原理"，像状语这种动词之前的非核心位置，也是比较常规的焦点重音的位置；因此，这里所谓的核心词移位的韵律动机并不一定存在。

鉴于上述的种种困难，我们怀疑：从句法操作上来解释"宾语代入"现象的路子，可能在根本上是错误的，至少是极不经济的。如果承认在词库和句法表达之间有论元结构这一表达层面，那么我们可以假设部分动词的论元结构在这一前句法(pre-syntax)层面上可以在某种语义、语用因素的驱动下发生变化，即产生新的、有标记的论

---

① 这一点是徐杰学长在"汉语话题和焦点学术讨论会"(2000年6月香港理工大学)期间提醒我注意的，冯胜利(2000：174)也简略地提及。

② 冯胜利(2000：174)用的是树形图，我们为了节省篇幅而改成括号式。

元结构;这种有标记的论元结构最终投射成有标记的句法结构。还是拿“写毛笔”作例子,“写”的常规的论元结构包括:施事和受事两个强制性的论元角色,以及工具、方式、处所等可选性的论元角色;记作〔写:A,(I/M/L),__,P〕。这种常规的论元结构可以投射成无标记的句法结构,如:“小明用毛笔写字”等。但是,在现实交际中,人们有时想突出工具、方式、处所等可选性的论元角色,强调它们在某种情形下不只是环境性的论元角色,而是受到动词所表示的动作直接影响的角色;于是,强行把它们转变为受事性的论元角色,而原有的受事论元只能引退、淡出为隐性论元(implicit argument),不能在句法结构上投射出来。比如,为了突出“写”的工具论元“毛笔”的受动性,把它转变为受事论元,并把原有的受事论元“字”挤出动词“写”所激活的语义场景(scenes)的透视域(perspective)之外;形成了“写”的有标记的论元结构〔写:A,S,P(I)〕。这种有标记的论元结构可以投射成有标记的句法结构,如:“小明写毛笔”等。也就是说,把某些动词的环境性论元占据宾语位置,归结为这些动词在论元结构层面上发生了论元结构转变的过程;突出地表现为从可选的环境性论元变为强制性的域内论元,并且在语义上获得受动性的特征。例如:

(5) a. ? 小明一天到晚写大字,都写秃了好几支毛笔了。

b. ?? 小明一天到晚写大字,把毛笔都写秃了好几支了。

c. 小明一天到晚写毛笔,都写秃了好几支毛笔了。

d. 小明一天到晚写毛笔,把毛笔都写秃了好几支了。

(6) a. ? 张老师喜欢写板书,一堂课下来写得黑板满满当当的。

b. ?? 张老师喜欢写板书,一堂课下来把黑板写得满满当当的。

c. 张老师喜欢写黑板,一堂课下来写得黑板满满当当的。

b. 张老师喜欢写黑板,一堂课下来把黑板写得满满当当的。

(7) 我抽他的雪茄，坐他的沙发，睡他的老婆。(《百变神偷》)

(5—6)的 a 和 b 不如 c 和 d 来得连贯和通顺，尽管“写大字”已经隐含了工具“毛笔”、“写板书”已经隐含了处所“黑板”；但是，后续小句直接把这个前一小句中隐含的工具、处所直接处理成受事的表达方式还是使人觉得突兀。而(5—6)的 c 和 d 在前一小句中已经把“写”的工具“毛笔”和处所“黑板”受事化了，因此为它们进入后续的受事宾语句 c 和有标记的处置受事的句式 d 作了足够充分的铺垫。(7)是更为极端的例子，这是电影《百变神偷》中警察局长在法官家里说的一句话。因为法官总是为窃富济贫的神偷辩护，使神偷免予刑事责任；警察局长恨透了这个法官，跟法官的后任太太偷情，并坐在法官家的沙发上一边抽法官的雪茄，一边说了上面这句泄愤的话。其中，“抽他的雪茄”是常规的“动作—受事”关系，属于无标记的表达；而“坐他的沙发”是特殊的“动作—受事性处所”关系，“睡他的老婆”是特殊的“动作—受事性共事”关系，它们都是有标记的表达。在警察局长的心目中，在法官家的沙发上坐、跟法官的太太通奸，和抽法官的雪茄一样，都是一种直接的处置性行为，并间接地使法官受到影响。这些例子有力地说明，代入宾语在语义上的确具有受动性。

总之，这一切证明介于词库和句法表达之间应该有论元结构这一语言知识的表达层面。在这一层面上，动词性成分的论元结构在某些语义、语用因素的促动下可以发生变化，即派生出新的有标记的论元结构。最终，这种新的、有标记的论元结构会直接投射到句法表达层面上来，即实现为有标记的句法表达形式。也就是说，解决词库中动词的句法、语义特性与表层的句法表达式之间的不一致，并非一定要诉诸句法操作这种不太经济的手段。

7 结语：多层面互动的限度和原则

句法、形态(词汇)和音系的多层面互动，显然不是漫无边际的；而是有一定的限度，并且遵循一定的原则的。比如，Elizabeth Selkirk 等人发展的韵律结构理论(prosodic structure theory，简称

PST)发现：(i) 音步层级以上的韵律成分的边界是由句法域(syntactic domain)的边界投射(选择)的。典型的情况是：一个词汇 X_0 的边界投射为一个韵律(或音系)词的边界，一个 XP 的边界投射为一个音系短语的边界。至于音节和音步的边界，不是由句法投射的，而是在词的基础上组织起来的。(ii) 音系学只能有限度地利用(has limited access to)句法学，因为从形态—句法边界到韵律边界的投射是音系学和句法学之间唯一的界面(interface)。一旦投射完毕，音系学就是完全独立的。(iii) 韵律结构是受到严格阶层假设(strict layer hypothesis)限制的。这个假设规定：韵律成分只能划分成音节、音步、韵律词、音系短语等数目有限的范畴，这些范畴是按照层级关系组织起来的；于是，像音系短语就只能支配(dominate)韵律词，韵律词依次只能支配音步。[①] Duanmu (1995)通过对上海话和台湾闽南话的变调域的研究，指出他在以下几个方面同意 PST：(i) 形态-句法边界投射为音系边界。(ii) 音系学只能从句法学那儿接受(assume)有限的信息。像名词、动词、形容词等句法范畴和成分统制关系等信息是不能利用的。需要的只是 X-bar 结构这一层面，于是一个 X_0 投射接受复合词(compound)重音，一个 XP 投射接受短语重音。(iii) 一旦适当的边界业已投射成功，那么就不必求助于句法信息了。(iv) 音系域和形态—句法域的错误匹配，可以由随后的独立的节律处理过程(比如，重音冲突的消解)来解释……(p. 251)。

但是，句法学对音系信息的依赖和利用的限度就不那么清楚了。像冯胜利(2000)企图用韵律要求来解释汉语“把”字句、“被”字句、主题句等句子中宾语位置的移动，动词之后的介宾结构中的介词贴附在动词上，历史上介宾结构位置从动词后向动词前的转移、SOV 结构向 SVO 结构的转移、以及“被”字句和“把”字句的产生和发展等历史句法问题，就颇让人怀疑：韵律对句法的作用会有这么大吗？不管答案是肯定还是否定，音系因素对句法的影响的限度和原则是目前多层面互动研究中最紧迫的课题，我们期待着能跟上述的 PST 相应的韵律句法理论的诞生。

---

① 详见 Duanmu (1995：250)。

鸣谢：本文是为《词汇语法语音的相互关联——第二届肯特岗国际汉语语言学圆桌会议论文集》（徐杰编，北京语言大学出版社近期出版）写的代前言。本文是笔者在日本御茶之水女子大学任教时写作的，蒙茶大的相原茂教授、东京大学的柯理思教授、筑波大学的刘勋宁教授和美国密歇根大学的端木三教授提供多种资料，还承茶大中文研究室助手森中野枝小姐帮我向茶大英文研究室借资料，谨此一并致以诚挚的谢意。

参考文献

陈　平（1988）《句法分析：从美国结构主义学派到转换生成语法学派》，《外语教学与研究》第4期。收入陈平（1991）《现代语言学研究——理论、方法与事实》，第31—54页，重庆出版社。本文据此。

陈　平（1994）《试论汉语中三种句子成分与语义成分的配位原则》，《中国语文》第3期。

丁声树等（1961）《现代汉语语法讲话》，商务印书馆。

范继淹（1964/1983）《汉语语法结构的层次分析问题》，《语法研究和探索》第1辑，第57—84页，北京大学出版社。收入《范继淹语言学论文集》，第211—238页，语文出版社，1986年。本文据此。

冯胜利（1997）《汉语的韵律、词法与句法》，北京大学出版社。

冯胜利（2000）《汉语韵律句法学》，上海教育出版社。

李　荣（1983）《方言研究中的若干问题》，《方言》第2期，第81—91页。收入《吴语论丛》，第3—14页，上海教育出版社，1988年。

林　焘（1957）《现代汉语补语轻音现象反映的语法和语义问题》，《北京大学学报》第3期。收入林焘（2001）《林焘语言学论文集》，第1—22页，商务印书馆；本文据此。

林　焘（1962）《现代汉语轻音和句法结构的关系》，《中国语文》7月号。收入林焘（2001）《林焘语言学论文集》，第23—48页，商务印书馆；本文据此。

吕叔湘（1942）《中国文法要略》（上卷），商务印书馆；中、下卷出版于1944年。现据《汉语语法丛书》，商务印书馆，1982年。

吕叔湘（1963）《现代汉语单双音节问题初探》，《中国语文》第1期，第10—22页。收入吕叔湘（1984）《汉语语法论文集》增订本，商务印书馆。

陆丙甫（1989）《结构、节奏、松紧、轻重在汉语中的相互作用》，《汉语学习》第3

期。
沈家煊（1999）《不对称和标记论》，江西教育出版社。
汤延池（1983）《国语语法的主要论题：兼评李讷与汤逊著〈汉语语法〉》（之一至之五），原文之一至之四刊载于《师大学报》1983年第二十八期，第391—441页；之五刊载于《中国语文》1984年第五五卷第二期，第22—28页。收入汤延池（1988）《汉语词法句法论集》，第149—240页，台湾学生书局印行；本文据此。
王　力（1943）《中国现代语法》（上册），商务印书馆；下册出版于1944年。现据《汉语语法丛书》，商务印书馆，1985年。
吴为善（1986）《现代汉语三音节组合规律初探》，《汉语学习》第5期。
吴为善（1987）《1＋3＋1音段的语法结构分析》，《汉语学习》第3期。
吴为善（1989）《论汉语后置单音节的粘附性》，《汉语学习》第1期。
邢福义（1991）《汉语里宾语代入现象之观察》，《邢福义自选集》第155—173页，河南教育出版社。
徐烈炯（1988）《生成语法理论》，上海外语教育出版社。
徐烈炯、刘丹青（1998）《话题的结构与功能》，上海教育出版社。
袁毓林（1998）《汉语动词的配价研究》，江西教育出版社。
袁毓林（2002a）《论元结构和句子结构互动的动因、机制和条件》，提交The 2nd Kent Ridge International Roundtable Conference on Chinese Linguistics（theme：Syntax-Morphology-Phonology Interface），National University of Singapore，November 27—29，2002.
袁毓林（2002b）《语言学中的"假设—演绎"法及其使用限制》，提交"岳麓论坛"·语言学研究方法论讨论会，长沙，湖南大学。将刊于《汉语学刊》2004年第2期。
张国宪（1989）《"动＋名"结构中单双音节动作动词功能差异初探》，《中国语文》第3期。
《中国语文》编辑部（1963）《语言学资料》6：描写语言学（语法部分）专号。
朱德熙（1982）《语法讲义》，商务印书馆。
朱德熙（1985）《语法答问》，商务印书馆。
Bao，Zhiming（1999）*The Structure of Tone*. Oxford：Oxford University Press.
Bybee，Joan & Revere Perkins & William Pagliuca（1994）*The Evolution of Grammar*：*Tense*，*Aspect*，*and Modality of the World*. Chicago：The University of Chicago Press.

Bloomfield, Leonard (1935) *Language*. New York: Allen & Unwin Ltd.

Chao, Yuen-Ren (1968) *A Grammar of Spoken Chinese*. University of California Press, Berkeley. 丁邦新 全译本《中国话的文法》,香港中文大学出版社,1980年,据刘梦溪主编《中国现代学术经典·赵元任卷》,胡明扬、王启龙编校,河北教育出版社,1996年。

Chen, Mathew Y. (2000) *Tone Sandhi: Patterns Across Chinese Dialects*. Cambridge Mass: Cambridge University Press.

Chomsky, Noam (1957) *Syntactic Structure*. The Hague: Mouton.《句法结构》,邢公畹、庞秉均、黄长著、林书武译,中国社会科学出版社,1979年。

Chomsky, Noam (1965) *Aspects of the Theory of Syntax*. Massachusetts: The MIT Press.《句法理论的若干问题》,黄长著、林书武、沈家煊译,中国社会科学出版社,1985年。

Chomsky, Noam (1982) *Lectures on Government and Binding. The Pisa Lectures*. Holland: Foris Publication.《支配与约束论集——比萨学术演讲》,周流溪、林书武、沈家煊译,赵世开校,中国社会科学出版社,1993年。

Chomsky, Noam (1995) *The Minimalist Program*. The MIT Press.

Cinque, Guglielmo (1993) A Null Theory of Phrase and Compound Stress. *Linguistic Inquiry*, 24, p. 239—297.

Crystal, David (1997) *A Dictionary of Linguistics and Phonetics*. Blackwell Publishers Ltd.《现代语言学词典》,沈家煊译,商务印书馆,2000年。

De Saussure (1961) *Course in General Linguistics*. W. Baskin (trans.), Peter Owen.《普通语言学教程》,高名凯译,商务印书馆,1981年。

Duanmu, San (1990) *A Formal Study of Syllable, Tone, Stress and Domain in Chinese Languages*. Ph. D. dissertation, Massachusetts Institute of Technology.

Duanmu, San (1995) Metrical and Tonal Phonology of Compounds in Two Chinese Dialects. *Language*, 71, p. 225—259.

Duanmu, San and Bing-fu Lu (1990) Word Length Variations in Chinese Two-word Constructions. Ms., MIT and University of Connecticut.

Duanmu, San (2000) Stress in Chinese. In Xu Debao (2000) (ed) *Chinese Phonology in Generative Grammar*, Academic Press, p. 117—138.

Goldberg, E. Adele (1995) *Constructions: A Construction Grammar Approach to Argument Structure*. Chicago and London: The University of Chica-

go Press.

Goldsmith, John (1976) *Autosegmental Phonology*, Ph. D. dissertation, MIT. [New York: Garland, 1979.]

Grimshaw, J. (1990) *Argument Structure*. MIT Press.

Hale, K. & S. J. Keyser (1991) *On the Syntax of Argument Structure*. Cambridge Mass: MIT Press.

Harris, S. Zellig (1946) From Morpheme to Utterance. *Language*, 22, p. 161—183; in Joos (1958) (ed), p. 142—153.《从语素到话语》,李振麟译,收入《中国语文》编辑部(1963)第14—28页。

Harris, S. Zellig (1951) *Methods in Structural Linguistics*. Chicago: The University of Chicago Press.

Heine, Bernd, Ulrike Claudi and Friederike Hunnemeyer (1991) *Grammaticalization: A Conceptual Framework*, Chicago: The University of Chicago Press.

Hockett, F. Charles (1942) A System of Descriptive Phonology. *Language*, 18, p. 3—21; in Joos (1958) (ed), p. 97—108.

Hockett, F. Charles (1947) Problems of Morphemic Analysis. *Language*, 23, p. 321—343; in Joos (1958) (ed), p. 229—242.《语素分析的一些问题》,程雨民 译,收入编辑部(1963)第54—67页。

Hockett, F. Charles (1954) Two Models of Grammatical Description. *Word*, 10, p. 210—230; in Joos (1958) (ed), p. 386—399.《语法分析的两种模型》,范继淹译,收入编辑部(1963),第96—110页;又收入《范继淹语言学论文集》,第309—347页,语文出版社,1986年。

Hockett, F. Charles (1955) *A Manual of Phonology*. Baltimore: Waverly Press.

Hockett, F. Charles (1958) *A Course in Modern Linguistics*. Macmillan Publishing Co., INC.《现代语言学教程》,索振羽、叶蜚声译,北京大学出版社。

Hopper, Paul and Elizabeth Traugott (1993) *Grammaticalization*. Cambridge: Cambridge University Press.

Jackendoff, Ray (1990) *Semantic Structure*. Cambridge Mass.: The MIT Press.

Jesperson, Otto (1933) *Essentials of English Grammar*, London: George Allen & Uwin Ltd.

Joos, M. (1958) (ed) *Readings in Linguistics: the Development of Descriptive Linguistics in America since 1925*. American Council of Learned Societies.

Levin, Beth & Malka Rappaport (1995) *Unaccusativity: At the Syntax-Lexical Interface*. Cambridge Mass.: MIT Press.

Lu, Bing-fu and San Duanmu (1991) A Case Study of the Relation between Rhythm and Syntax in Chinese. Paper presented at the Third North America Conference on Chinese Linguistics, May 3—5, Ithaca.

Newmeyer, J. Frederick (1986) *Linguistic Theory in America*, Second Edition. Orlando: Academic Press, INC.《当代美国语言学史》,吴黉铭译,台北:文鹤出版有限公司,1998年。

Shih, Chi-lin (1986) *The Prosodic Domain of Tone Sandhi in Chinese*. Ph. D. dissertation, University of California, San Diego.

Wells, S. Rulon (1947) Immediate Constituents. *Language*, 23, p. 81—117; in Joos (1958) (ed), p. 186—207.《直接成分》赵世开 译,收入编辑部(1963)第29—53页。

Zwicky, Arnold (1977) *On Clitics*, Bloomington: Indiana University Press.

2003年5月初稿,9月改定

(发表于《语言科学》2003年第6期)

五、附　　录



赵元任先生评传

1 语言奇才 学人生涯

赵元任(1892—1982),著名语言学家,中国现代语言学的开创者之一。1892年11月3日生于天津紫竹林,4岁开蒙,1902年(10岁)开始在家里读私塾,1906年(14岁)在老家常州接受新式学校教育,1907年(15岁)转入南京江南高等学堂预科。1910年(18岁)7月21日参加清华学校庚子赔款留学美国学生考试,以73.2/3分的成绩被录取为第二次考取清华学校庚子赔款留学美国学生榜第二名。8月下旬赴美入康奈尔大学,主修数学。1914年(22岁)毕业后继续在该校修习哲学。1915年(23岁)转入哈佛大学,1918年(26岁)获哲学博士学位。1919年(27岁)任康奈尔大学物理学讲师。1920年(28岁)回国到清华学校任教。1921—1924年赴美任哈佛大学哲学系讲师、教授。1925年(33岁)回国,1925—1929年任清华学校(国学)研究院导师。1929—1938年任中央研究院历史语言研究所研究员兼语言组主任。1938年(48岁)赴美,先后在夏威夷大学、耶鲁大学、哈佛大学、加利福尼亚大学伯克莱分校任教。1945年当选为美国语言学会会长。1981年被北京大学授予名誉教授称号。1982年病逝于美国麻省剑桥,终年九十岁。

赵元任从小对学习语言和方言有特别的兴趣,并表现出极高的天赋。小时候由于一家三代跟着祖父赵执诒在河北各地居住,因而起先说的不是家乡话常州话,而是一种南方口音很重的北京话;比如,舌尖后辅音声母(即卷舌音zh-,ch-,shi-,r-)混同于舌尖前和舌尖中辅音声母(即z-,c-,s-,l-),分不清前鼻音韵尾-n和后鼻音韵尾-ng。后来,又从带他的周妈那儿学了保定话。赵元任学的第一种别处的话不是他本乡的常州话,而是江苏常熟话;因为远嫁江苏常熟

的姑妈带着两个表弟来北方省亲，为了跟他们玩耍，在5岁那年他就学会了说一种地道的常熟话，尽管当时他念书用的是常州音。1901年（9岁）祖父去世，他随家人回常州居住和上学，才开始学常州话。1904年（12岁）那年父母亲同年病逝，第二年他到苏州大姨娘庞冯氏家居住；于是，他自然地学会了第四种方言苏州话。后来，他又从在福州住过多年的伯母那儿学了一点儿福州话。1907年（15岁）去南京读书，他尽管不喜欢南京的口音，但是出于对方言的特别兴趣，不久就学会了南京话。在南京的三年中，跟室友福州人邵绳武交换方言，互相教常州话和福州话；最终，他又学到了更多的福州话。到南京的第二年（1908年），他开始跟美国先生嘉化（David John Carver）在课堂上学英语。1910年（18岁），在南京三年预科没读完，进京投考官费留美生；在北京预备了一个春天，期间还学了一阵子拉丁文，这也是当选科之一。进了康奈尔大学以后，他又学了德语；他还从一个国际函授学校学了法语。在此期间，因为跟无锡人胡明复同住，就学会了无锡话。进哈佛大学研究院以后，他还选学了梵文。回国以后，他又学会了上海话、湖北话等汉语方言。1920年英国哲学家罗素（Bertrand Russell）访华作学术演讲，赵元任担任翻译。有一次，他陪罗素坐长江轮船到长沙演讲，向同船的邀请罗素的主人（长沙人）学了湖南话。到了长沙以后，他居然用那种国语底子的湖南话作翻译。赵元任的语言天才和丰富的语言和方言经验，为他从事语言研究提供了得天独厚的条件。

赵元任从事语言学理论和汉语语言学研究长达60余年，在推行国语、设计汉语拼音方案、汉语方言调查和研究、音位学理论研究和记音方法与记音工具的设计、汉语语法研究、普通语言学理论研究等方面都有重要的贡献。他共出版著作10余种，论文60余篇。赵元任六、七十年代的论文或讲演稿由 Anwar S. Dil 编成 *Aspects of Chinese Sociolinguistics, Essays by Yuen Ren Chao*（汉语社会语言学面面观-赵元任论文选，Stanford University Press，1976）。叶蜚声选译了其中的3篇，加上30年代的一篇，编成《赵元任语言学论文选》（伍铁平校，中国社会科学出版社，1985年）。为了纪念赵元任诞辰100周年，袁毓林选了赵元任的12篇论文，编成《中国现代语言学

的开拓和发展-赵元任语言学论文选》(清华大学出版社,1992年)。

2 推广国语 设计拼音

20年代初,赵元任先生回国以后不久,就满怀热情地投身于当时的国语统一运动。他1920年8月回国,9月18日就参加了在北京举行的国语统一运动筹备委员会会议。他不仅编写国语课本、灌制配套唱片;还根据语音学原理设计了拉丁字母式的汉语拼音方案,为汉语书写系统的拼音化和拉丁化奠定了坚实的基础。

赵元任1922年由商务印书馆出版《国语留声片课本》,还发行了由他发音的配套唱片。所依据的标准音是《国音字典》(1920)中兼顾古今南北的老国音。这种老国音与京音(北京音)稍有差别,比如:京音没有入声,国音有入声;京音不分尖团,国音分尖团。后来,据赵元任自己说,只剩下他一个人会说这种有入声的老国音。事实证明,这种兼顾古今南北的国音是无法推广的。在哈佛大学哲学系任教期间,赵元任1922年在《国语周刊》1卷7期上发表《国语罗马字的研究》。1923年,又在《中国留学生月刊》(*The Chinese Students' Monthly*)18卷8期上发表文章"Principles of Romanization",提出了建立实用的国语罗马字系统应该考虑的25条原则。这为汉语书写系统的拼音化和拉丁化作出了充分的理论准备。

赵元任1925年回清华(国学)研究院任教不久,于1925年9月参加由刘复(半农)发起的"数人会";与刘复、钱玄同、黎锦熙、汪怡、林语堂、周辨明诸君子讨论国语运动问题。他们经过一年时间,开会22次,九易其稿,终于拟定《国语罗马字拼音法式》稿本。1926年9月,"国语统一筹备会"召开"国语罗马字拼音研究委员会",通过并提请教育部公布。1928年9月26日,国民政府大学院公布这套由赵元任设计的国语罗马字方案,作为"国音字母第二式"。他还用它来翻译《最后五分钟》。1935年赵元任由商务印书馆出版《新国语留声片课本》甲种(注音符号本)和乙种(国语罗马字本),还发行了由他发音的唱片16面。所依据的标准音是《国音常用字汇》(1932)中紧靠北京话的新国音,把老国音中北京话所没有的音都取消了;因此,这

跟中华人民共和国建立后以北京话为标准音的普通话是一致的。到了晚年，他老骥伏枥，孜孜不倦地研究“通字”。所谓“通字”就是在国语罗马字的基础上，增加吴语、粤语、闽语的区别。1973 年，他把《通字草案》带回国，听取意见，继续修改。1983 年，商务印书馆编辑部把它译成中文，以中英对照的方式出版。

赵元任发扬中国传统语言学注重实践的优良传统，学以致用、身体力行，积极参加国语统一运动、亲手设计国语罗马字，从理论到实践，在推行国语工作和设计汉语拼音方案方面起了示范作用。

3 调查方言 建立新学

本世纪 20 年代是世界语言学发生深刻变化的时期，当时有两股研究思潮推动着现代语言学的诞生。一股是传统语言学阵营中的结构主义语言学思潮，瑞士语言学家索绪尔(F. de Saussure, 1857—1913)在长期的印欧系古代语言和比较语言学的研究实践中逐步形成了对语言的符号性质和系统价值的独到认识，首创了普通语言学这一学科。其《普通语言学教程》(1916)由其学生整理出版后，在语言学界产生了广泛的影响，使 19 世纪以历时为主的语言研究转变为 20 世纪以共时为主的结构语言学。另一股是美国人类学阵营中的描写语言学思潮，以鲍阿斯(F. Boas)为代表的人类学家在调查、记录美国的土著语言(印第安人的各种语言)的实践中，逐步形成了一套处理陌生语言的分析方法。其后萨丕尔(E. Sapir)出版了《语言》(1921)，布龙菲尔德(L. Bloomfield)出版了《语言论》(1933)，发展出一套严格的语言单位的发现程序(discovery procedure)。

赵元任两度留美，自然受美国描写语言学的影响较大，加上中国学术注重实践的传统，所以他 1925 年应聘到清华(国学)研究院任导师后，立即开始了他一生中最重要的学术活动——汉语方言调查。他带着杨时逢赴江苏、浙江实地进行吴方言的调查，先写出论文《北京、苏州、常州语助词的研究》(刊《清华学报》1926 年第 3 卷第 2 期)。而后写出专著《现代吴语的研究》(清华学校研究院 1928 年印行)，这是中国第一部用现代语言学方法研究方言的著作，赵元任也

因此获得了中国现代语言学的奠基人的称誉。

《现代吴语的研究》共6章，分"吴音"和"吴语"两部分。前4章是"吴音"，讨论各处吴语的声母、韵母和声调的音类和音值，列举各地的语音特点，总结吴语的共同特征，指出"吴语为江苏浙江当中'并定群'等母带音，或不带音而有带音气流的语言"(第88页)。后2章是"吴语"，讨论词汇、语法等问题，主要内容是30个方言点75个词的词汇对照表和22个方言点56种用法的语助词对照表。词汇对照表之后列举各地特别的词，如上海话"白相"(玩儿)、温州话"吃天光"(吃早饭)等。语助词对照表之后有成篇的记音材料。此外，书后还附有作者调查时所用的各种表格。赵元任在方言调查时最先使用国际音标记录汉语方言，[1]语音分析深入细致，并能联系古代音韵考察汉语的古今变化，使错综复杂的语言现象得到科学合理的解释。各地声韵调和词语异同都用表格形式表示，便于对照比较。由于该书调查点偏重江苏，浙江中部和南部的调查点较少，因而吴方言的复杂情况还没有得到充分的反映。[2] 尽管如此，赵元任对吴语界限的规定今天看来还是很合理的，他在书中(第1页)说："广义的吴语包括江南的东南部跟浙江东北大半部。这吴语观念的定义或这观念的能否成立是要等详细研究过后才知道，现在暂定的'工作的假设'就是暂以帮滂并，端透定，见溪群三级分法为吴语的特征。"因为吴方言塞音声母有浊塞音[b, d, g]、不送气清塞音[p, t, k]和送气清塞音(p', t', k')三套，而官话和大部分南方方言只有后两套；所以浊音的有无可以作为划分吴方言和其他方言界线的一条关键的同语线(Isoglos)。现代的各种方言地图就是根据这一同语线来划分吴方言和其他方言的分界的。

---

① 高本汉早年调查汉语方言用的是其师 J. A. Lundell 所创造的瑞典方言字母，见高本汉(1940)第142页。

② 参考《中国大百科全书·语言文字》，第421页。

4 借助国学 创制字表

赵元任用描写语言学的方法调查汉语方言，同时又充分利用中国古代的音韵学知识来控制方言调查。具体的做法是选择比较常用的3,567个单字，按照《切韵》《广韵》一系韵书所代表的中古音系统排列成表，形成一整套的“方音调查表格”。表格中的字先按13摄排列(假摄归并入果摄、江摄归并入宕摄、曾摄归并入梗摄)，同摄的先分开口合口，再分一二三四等。相承的四声并列，每页横行的韵目举平以赅上去，竖行按36字母排列。这样，声母、韵母、声调搭成框架，每个字放在各自应占有的音韵地位，形成一张张韵图，相当于现代汉语声韵调配合表。用这种字表调查方言的音系，不但便于归纳整理出所调查的方言的音系，而且使许多复杂不易解释的现象大都可以得到理解，从而得出方言音系在古今语音演变方面的条理。

后来，中央研究院历史语言研究所正式刊印了赵元任制的《方音调查表格》(1930)。中华人民共和国成立后，中国科学院语言研究所在这本《表格》的基础上，删去原表格中不必要的罗马字注音，①还删去了一些不常用的字和又音字，改正了一些字的音韵地位，加入了一些常用字，删改和增补了一些字的注释，改编成《方言调查字表》(1957)。几十年的实践证明，用这种字表作为调查汉语方言语音的基础是一个比较简便易行的办法，同时它还可以作为学习和研究汉语音韵的参考资料。② 这也是中国的描写语言学一开始就不同于美国描写语言学的一个重要的方面。原来，中国的描写语言学一直是拿历史语言学作背景的；甚至可以这么说，中国的方言研究是共时描写和历时研究结合得最好的领域。

① 这使人想起当时的一句笑话，“胡适之无往而不注，赵元任无往而不音”。

② 参考《中国大百科全书·语言文字》，第79页。

5　利用新学　更新旧学

在赵元任的影响下，20～40 年代中国的方言调查呈现出活跃的气象，出版了一批具有较高学术价值的方言研究著作。比如，赵元任的《钟祥方言记》(1939)和《中山方言》(1948)、罗常培的《厦门音系》(1930,)和《临川音系》(1940)、赵元任等的《湖北方言调查报告》(1948)。这种方言调查研究的重要性，只有联系本世纪初(五四前后)汉语音韵学研究出现了严重危机的历史背景才能充分体现出来。大家知道，汉语具有悠久的历史，从先秦到现代汉语的语音发生了重大的变化。但是，对这种古今语音演变的研究一直是很不充分的。其中一个重要的原因是，汉语是用表意的方块汉字来记录的，方块汉字不跟统一的语音相联系，表现出超时代、跨地域的特点；同一个汉字，不同时代、不同地区的人虽然写法相同、理解相同，但读音可以完全不同。因此，后人无法从文字上了解古代的语音面貌和历史流变。于是，传统的音韵学研究只能根据各个时期的诗韵以及韵书、韵图等书面文献，来归纳各个时期的音类，却无法知道每一个音类的具体音值。比如，由顾炎武开始的清代古音学利用《诗经》的用韵和汉字的谐声分析先秦的古音，归纳出《诗经》的韵部，整理出谐声的系列，弄清了从上古到中古的韵类分合的演变情况，但无法对它进行具体的语音学描写。关于《切韵》系统的韵书的研究，音韵学家们以反切为主、以韵图为辅，把中古时期的音类大致分别清楚，但又无法推测它们的读法。[①] 就这样，囿于纸上的材料，造成了音韵学研究的危机。所幸的是，现代汉语的各种方言都是从古代汉语演变来的，根据一般语言研究的结果，语音的演变大致都有途径可寻，方言的差异就是演变途径不同的结果。[②] 所以，综合考察南北方言的差异，可以窥见古音读法的大概面貌。在这方面，瑞典汉学家高本汉(Bernhard Karlgren,1889 — 1978)捷足先登，他凭恃良好的历史比较语言学修养和

① 详见徐通锵(1991)，第 4、44、124—125 页。

② 参考董同龢(1974)，第 139 页。

精湛的瑞典方言调查经验，利用汉语的33处方言(包括日译吴音、汉音，高丽译音，安南译音等境外方言)，结合《切韵》等“历史上的旧材料”，给以《切韵》为代表的中古音类——拟测(reconstruct)了具体的音值，[①]为中古音的研究奠定了新的基础；使汉语音韵学除了分类之外，在拟音上有了一套合适的方法和方便的工具。

由于引进方言材料作为考订古音的佐证，使濒于危机的汉语音韵学获得了新生，也使人们重新认识到方言资料在汉语史研究中的价值。赵元任于1921年得到高本汉《中国音韵学研究》的前三册，[②]对方言调查和研究的重要性是看得很清楚的。这就难怪他一回清华后，马上要一头扎进方言调查中去了。他和罗常培、李方桂历时四五年把高氏巨著移译过来，这对汉语音韵学研究的推动作用也是不可低估的。

6 洞烛幽微 发明胜义

赵元任在对吴语的调查研究中，形成了对吴语的性质和源头的独到认识。从表面上看，《现代吴语的研究》是一部现代吴方言的描写比较语音学著作，但实际上已隐含着方言差异的比较和重建原始吴方言之间的内在联系。书中说，“所用材料范围甚小，对吴语的事实虽多所发现，而对于空间与时间上的远处的推测没有什么发明”(第3页)。这些话暗示可以通过方言差异的比较去作“空间与时间上的远处的推测”。有些语言学家从中得到启发，设法通过方言内部的差异的比较去重建原始方言。比如，罗杰瑞(美国华盛顿大学东亚系教授)受赵元任这些话的启发，提出了“原始闽语”这个设想，并对原始闽语作了全面的拟测。[③]

指出原始方言这个事实是很重要的，它将使汉语语音史研究的理论框架发生变革。高本汉虽然认识到“在近古汉语的时候已有不

---

① 详见高本汉(1940)，原著于1915、1916、1919、1926分四册出版。

② 见高本汉(1940)，第7页，译者序。

③ 详见徐通锵(1991)，第144—145页。

同的方言”(第238页),但他还是坚信“在主要特征上我们这部书所研究的每一种方言都成一种从《切韵》所代表的古代汉语直接演变下来的缩影”,因为“开口合口,颚介音成素、各摄的主要元音,所有《切韵》语言的特征,在我们的方言里大体上都有完全合乎规律的对映”(第528页)。因此,高本汉认为《切韵》代表7世纪长安音系,它是现代汉语各方言的原始母语,可以“把一切方言都跟《切韵》的语言连接起来”(第528页),用《切韵》来解释现代方言的歧异。现在,由赵元任隐含、由罗杰瑞发明的原始方言的观念,对汉语语音史的研究将产生某种反拨的作用。比如,张琨(美国加州大学伯克利分校东亚系教授)充分认识到《切韵》代表一种综合音系,反映“南北是非,古今通塞”的特点。指出既然《切韵》包含各地方言的特点,那么在研究它与现代汉语各方言的关系时,就不应该笼统地把它看成母语,而需要根据不同方言的特点,把《切韵》的音类加以简化,剔除一些不属于该方言的因素,建立原始方言的音系,进而比较各个原始方言之间的差异去建立原始汉语。[①] 这真可谓一句话引出了一连串研究和真知灼见。

赵元任在调查方言时,往往能注意到一些特殊的现象,并敏锐地发现其中所隐含的理论价值,尝试用以解决一些重大的音韵学问题。比如,他调查、分析了民间8种利用反切方式构成的秘密语,发现在反切语中有介音属声属韵的问题;有些反切语的i介音两属,即反切上下字都要求有i介音;有些反切语的i介音属声,有些则属韵。[②] 从而提出了“介音和谐”说,用以解决关于j化声母的争论。原来,高本汉根据反切上字分组的趋势(一二四等是一组,三等是一组),提出了三等字声母j化的观点。[③] 事实上,这种分组是不严格的,各有拿对方的字作反切上字;并且,精清从心四母的反切上字也有分组的趋势,但高氏为了顾全自己的体系,并没有把它们分为两类(因为他已经把章、昌、船、书4母拟测为ts’(t’),就不能再把精组各母分为单

① 详见徐通锵(1991),第141—142页。

② 详见赵元任(1931)《反切语八种》,中央研究院《历史语言研究所集刊》二本三分。

③ 详见高本汉(1940),第28—30页。

纯的 ts 和 j 化的 tsj 两类了)。这引起了学术界的争论和批评。对此,赵元任的见解是:“关于高本汉的纯声母和 j 化声母,我们用介音和谐的概念来代替 j 化的概念。原则是这样的,韵母以闭 i 开始的字,它的反切上字的韵母趋于以闭 i 开始,韵母以开 i 或其他元音开始的字,它的反切上字的韵母也趋于以开 i 或其他元音开始”。[①] 后来,李荣(1952)在此基础上进一步作研究,终于使这一点成为定论:高本汉把《切韵》的声母分成单纯和 j 化两类是没有根据的。[②]

7 精研语法 功盖后世

1948 年,赵元任先生出版《国语入门》(*Mandarin Primer*, Harvard University Press)。后来,李荣先生把其中跟语法有关的部分编译为《北京口语语法》(开明书店,1952 年)。这称得上是中国第一部尝试运用结构主义语言学的方法研究汉语语法的著作,在语法分析的理论、方法和体系上,对以后的汉语语法研究产生了极为深远的影响。

赵元任 1965 年试版(pre-edition)、1968 年正式出版《中国话的文法》(*A Grammar of Spoken Chinese*, University of California Press)。该书采用美国描写语言学的理论和方法来全面而系统地描写和分析现代汉语语法,材料丰富、方法严谨、论述精到、体系分明。全书分八章,依次为:序论、句子、词和语素、形态类型、句法类型、复合词、词类和体词、动词和其他词类。基本上重词法、轻句法。在语法分析的方法上,该书以直接成分分析法作为分析语法结构的主要方法,使语言结构的层次观念在汉语语法学界更加深入人心。并且,使得诸如直接成分、开放类和封闭类、自由形式和黏着形式、向心结构和离心结构、结合面的宽或窄等,成为汉语语法描写不可缺少的内容。在语言材料的选取方面,该书主要依据北京方言,而且是大量采

---

① 见赵元任(1941)Distinctions within Ancient Chinese,刊 *Harvard Journal of Asiatic Studies*(哈佛燕京学报),第五卷第 2 期,第 214 页。

② 详见徐通锵(1991),第 132—134 页;李荣(1956),第 107—110 页。

用非正式的日常口语;这跟过去注重古代汉语或现代书面语的传统很不一样,更多地体现了美国描写语言学派的精神。正因为赵元任注重口语,因而他在语法分析中就特别重视语音特征,关注语法和语音、节律的关系。该书除了开头专辟一节讲汉语语音外,在讨论词、结构、复合句等的定义时引入语音特征,随时讨论停顿、轻重音、升降调对语法的影响。在讨论的时候,不断地拿其他汉语方言跟北京话作比较,拿英语、德语等外语跟汉语作比较,颇具理论语言学的特色。该书一方面注重不同语言的共性,同时不抹杀汉语语法的特点;提出了汉语的主语和谓语是话题和说明的观点、句子可以有大主语和小主语、一个整句可以由两个零句构成、动词可以作主语等观点,对后来的汉语语法研究产生了深远而重要的影响。在词类划分上,该书根据语法功能来分别词类;基本的出发点是:“语法描写的很大一部分是语言形式的分类”(p.2),“语法是研究一类一类的形式出现或不出现在由别的类构成的框架或槽之中的”(p.5);明确地指出词类划分的原则是功能,说“形式类是语言形式按其功能分的类,……一个词类是一个其成员都是词的形式类”(p.496)。他坚定地按分布了给词语分类,根据分布为每一类词下了严格的定义,这种做法跟以前的汉语语法著作有很大的不同。在具体分类上,他把体词分为名词、专名、处所词、时间词、“限定词+量词”复合词(三斤、这回)、“名词+定位词”复合词(墙上、饭前)、限定词(三、每)、量词、定位词(里、上)、代名词(我、什么)。对于量词、助动词、介词等封闭类,尽可能穷尽地列举其成员,一一描写它们的功能和用法。这些也都是以往所没有的。尤其是对语助词的分析,观察细致入微、描写准确周到。值得一提的是,赵元任对待各种理论、方法和语言现象,总是持一种开放、通达、适度的态度。比如,在语法分析时,他接受结构主义的思想,注重对语言的形式分析;但是,“并不取消意义的用处”(p.7),在说明一个词语表达的意义时经常联系词语使用的环境。对一句话能不能说,经常不作绝对判断;而是注明在什么场合下什么样的人这么说。他承认许多语法现象是个程度问题或频率问题。他虽然接受结构主义的理论,但是立论通达,从不拿事实迁就理论。他说:“在语言现象中寻找系统性和对称性在方法学上是可取的,只要不走得太远”(p.9—

10)。因此,他在寻找语法的系统性和对称性的同时,又注意不对称的一面和扭曲关系,充分尊重语言事实。作为例证,他指出:虽然在真正的动宾结构中,重音总是落在宾语上;可是,重音在第二音节上,并不一定都是动宾结构。比如,"烙'饼、炒'饭"在语法上是两可的,动词可以拿名词作宾语,也可以修饰它。这就是扭曲关系,即一种有时规则和对称、有时不规则和不对称的现象。

总的来说,《中国话的文法》系统地运用结构主义语言学的方法,对汉语语法事实进行了全面的描写和精到的分析,使得该书至今仍是国内外引用最多的汉语语法著作。甚至可以说,无论从涉及的语法事实的广度、分析的深入和细腻,还是从理论和方法的建树上,将近40年后的今天,还没有一部汉语语法著作能全面地超过它。

8 提升理论 回馈世界

在长期的汉语方言调查的基础上,赵元任对音位学理论进行了深入的研究,写出了《音位标音法的多能性》(1934)。文章阐明从语音材料归纳音位系统时可以有多种选择,答案不是唯一的。而影响答案的因素有:(1) 音位的尺寸问题,比如把塞擦音[tʃ, dʒ, tɕ]等看作一个音位可能是分析不足,而看作两个音位(塞音和擦音)可能是分析过头;(2) 组类问题,把哪些音归纳为一个音位会受到下列因素的影响:(a) 音质的准确度,(b) 系统的简单或对称的要求,(c) 本地人对于音类的见解,(d) 字源的顾及……,(3) 符号的选择,由于语音符号的使用有不少互相冲突的传统,因而音位归纳时常常要放弃其他方面的考虑来迁就现成的某套符号。[①] 文章立论通达,用例恰切,成为音位学理论的经典文献,一直为各国语言学家广泛引用。美国语言学家裘斯(M,Joos)在 *Readings in Linguistics: the Development of Descriptive Linguistics in America since 1925*(《语言学选读——1925年以来美国描写语言学的发展》,American Council of

① 详见赵元任(1934) The Non-uniqueness of Phonemic Solution of Phonetic Systems,中央研究院《历史语言研究所集刊》四本四分。

Learned Societies，1957)中收录了此文，并作了简短的评论，其中说到“我们很难想到有比赵元任的这篇文章更好的对早期音位学具有指导意义的单篇论文了”。

值得一提的是，赵元任在语音分析的实践中，凭着自己的声学和音乐修养，精心设计了一套五度制的标调字母，[①]为记录和研究汉语以及世界上其他有声调的语言提供了准确、方便的工具，为世界语言学界普遍采用。

1959 年，2 月 2 日至 4 月 1 日，赵元任应邀到台湾大学文学院中文系作“语言问题”系列演讲(16 讲)，并由台湾大学文学院出版专著《语言问题》。后来，他把这部书改写成英文本的 *Language and Symbolic Systems*(《语言和符号系统》，Cambridge University Press，1968)。《语言问题》于 1980 年又由商务印书馆出了新版。《语言问题》是他系统地讲述语言学以及相关的问题的演讲记录。他用风趣的语言、丰富而生动的例子阐明深刻的见解，先后被翻译为法语、西班牙语、葡萄牙语和日语。

赵元任能取得这么大的成就，固然与他天资聪慧、工作勤奋有关，但更主要的一点是赵先生具有极为广阔的学术背景，可以概括为：融会古今、贯通中外、横跨文理、精通音乐。赵先生从幼年开始诵习四书五经，对许多中国古代典籍烂熟于心，这一点只消看一下赵先生著作中信手拈来的引经据典就足够了。赵元任年轻时听过著名语言学家 J. Vendryes、Daniel Johns 和汉学家伯希和、马伯乐等人的课，还与美国描写语言学派的代表人物 Edward Sapir、Leonard Bloomfield、Bernard Bloch、Charles F. Hockett 等人讨论过语言学问题，对中外学术思想有深刻的领会。加上赵先生游历广泛，走到哪儿学哪儿的话，并且学哪儿的话象哪儿的话。先后学会了英语、法语、德语等多种语言和汉语的北京话、常州话、苏州话、常熟话、南京话、福州话、上海话、无锡话、湖北话……乃至“中国主要的方言系中每一系都会说一种”，丰富的语言经验为赵先生的语言研究提供了不

---

① 详见赵元任(1930) A System of Tone Letters，*Le Maître Phonélique*，troisième série，no，30，Avril-Juin。

尽的源头活水。清华大学早年倡导兼通古今中外的学术风格，与赵元任同时作清华(国学)研究院导师的梁启超、王国维、陈寅恪诸公，莫不做到古今中外融会贯通。在赵先生身上，还多出横跨文理这一层特色。赵元任先生上大学时专修数学，攻读博士学位时专修哲学；博士论文是关于数理逻辑和方法论的，题目是 Continuity: A Study of Methodology(《连续性——方法论的研究》)。毕业后，他在康奈尔大学教过物理，对声学方面特别感兴趣。良好的数理修养着实为赵先生从事语音的实验研究提供了利器，使得赵先生能掌握技术性很强的信息论，并能从信息论的立场来分析语言现象。更为重要的是，文理兼通的知识结构，使赵先生能够很及时地汲取当代自然科学的理论营养，形成新型的思维方式。比如，在《说清浊》中，赵先生主张清浊只用于指声母是不带音(清)或带音(浊)的，因为这种用法符合人的音感：不带音频率高，听起来觉得清；带音频率低，听起来觉得浊。但赵先生并不认为这种用法就是完美的名符其实，因为清音是噪音，频率带杂乱不清；浊音是乐音，频率带十分清楚。这样，从不带音和带音的声学特征上看，清与浊这对名称好像用颠倒了。怎么办呢？还是看看赵先生的解释：

> 可见音分 voiceless, voiced 并不是唯一的主要的发音方法的分别，以清浊的名词来配 Voiceless, Voiced 也只是为求逻辑上的整齐方便，也不是天经地义。大凡一种理论求其整齐紧凑就可能只照顾到事实的一部，一方面；如果求其包括的事实丰富，多方面来照顾，系统就不免会松弛下来。这也是丹麦的 Niels Bohr 教授常常讲的对补原则(principle of complementarity)。这本来是讲质子的动量与地位之间的相互关系，可是 Bohr 给它推广了用在好多问题上。

这就涉及到上世纪二三十年代发生在量子物理学界的一场观念革命：海森堡(Heisenberg)发现了测不准关系，大意是由于微观粒子具有波粒两象性，我们不能同时准确地测定粒子的位置和动量，两者总是存在着不确定性——如果决定粒子的坐标越准确，那么决定粒子在该坐标方向上的动量分量的准确度就越差；反之亦然。对此，玻尔(Bohr)创造了一种全新的逻辑工具叫做互补性(complementarity)。

互补性代表一些概念之间一种完全新型的逻辑关系：这些概念是互斥的，从而不能同时被考虑，因为那将导致逻辑上的错误；但是，为了对现象作出一种完备的描述，这些概念又全都是必需的。互补性引入了一种把我们的概念安排在里面的逻辑构架，意味着我们在谈论自然现象的方式上的一种巨大的扩充。量子力学在很大程度上改变了我们对世界的固有看法。赵元任又得风气之先，用这种新型的眼光看待语言学问题，高人一筹。赵元任是一个多方面的学者，深湛的音乐造诣又为赵先生精细的听音、辨音和声调、语调研究提供了得天独厚的助益。

从上面的介绍可以看出，赵元任的语言学研究具有贯通中西、融会古今的学术品格。一方面，积极地引进西方先进的现代语言学理论和方法，并深深地植根于汉语的土壤中；同时努力地利用中国传统学问中的各种材料，来建设富有时代气息和民族特色的中国现代语言学。另一方面，通过以现代语言学为指导的汉语方言调查和研究，为中国传统音韵学研究摆脱危机提供了丰富的新鲜材料，同时，在研究汉语的实践中不断地总结经验，并提升到普语言学理论的层面上进行概括，主动地让中国语言学回馈世界语言学。正因为这样，他的音位学理论使中国语言第一次对普通语言学产生影响。从而改变了中国语言学一向自立门户，游离于普通语言学之外的局面。现在，王士元（美国加州大学伯克利分校语言学系教授）等一批在美的华裔语言学家，通过研究汉语方言中的扩散性音变的各种复杂情况，提出了著名的词汇扩散理论（lexical diffusion theory），[①]在世界上的历史语言学界产生很大的影响，从而使中国语言学第二次对普通语言学作出杰出的贡献。

从中我们得到的启发是：只有真正贯通中西、融会古今的学术研究，才能走出国门、汇入世界学术的洪流。

鸣谢：本文为清华学校（国学）研究院成立70周年纪念会而作，写作时蒙张清常先生慨借赵元任制的《方音调查字表》（1930），谨此

---

① 详见王士元（1982）《语言变化的词汇透视》，《语言研究》第2期。

致以诚挚的谢意。

参考文献

董同和 (1974)《汉语音韵学》,台湾学生书局。

高本汉 (1940)《中国音韵学研究》,赵元任、罗常培、李方桂译,据台湾商务印书馆 1982 年版。

季羡林 (1988) 主编《中国大百科全书·语言文字》卷,"赵元任"、"国语罗马字"、"汉语方言字表"等条目,中国大百科全书出版社。

李士重 (1981)《〈汉语口语语法〉读后》,《中国语文》第 3 期。

李　荣 (1952)《切韵音系》,据科学出版社,1956 年版。

李　荣 (1982)《赵元任》,《方言》第 2 期;收入李荣(1985)《语文论衡》,商务印书馆。本文据此。

林　焘 (2002) 主编《20 世纪学术大典·语言学》卷,"赵元任"、"中国话的文法"等条目,福建教育出版社。

陆志韦 (1947)《古音说略》,重刊于《陆志韦语言学著作集》(一),中华书局,1985 年。

徐通锵 (1991)《历史语言学》,商务印书馆。

袁毓林 (1995)《融会古今 中西对流——赵元任早年的语言学研究及其影响》,《汉语学习》第 6 期。

赵新那 (1992a)《赵元任著作目录》,中南工业大学出版社。

赵新那 (1992b)《赵元任生平大事记》、《赵元任语言学论著要目》,见袁毓林编《中国现代语言学的开拓和发展-赵元任语言学论文选》的附录,清华大学出版社,1992 年。

赵元任 (1928)《现代吴语的研究》,据科学出版社,1956 年版。

赵元任 (1971)《我的语言自传》,《历史语言研究所集刊》,第 43 本第 3 分;收入赵元任(1985),第 87—106 页。本文据此。

赵元任 (1985)《赵元任语言学论文选》,叶蜚声译,伍铁平校,中国社会科学出版社。

赵元任 (1992)《中国现代语言学的开拓和发展——赵元任语言学论文选》,袁毓林编,清华大学出版社 1992 年。

(据《融会古今 中西对流——赵元任早年的语言学研究及其影响》改写,原载《汉语学习》1995 年第 6 期)

朱德熙先生评传

1　坚忍好学的人生　卓越辉煌的成就[①]

朱德熙先生，1920年10月24日生，江苏省苏州人氏。少年时代，他在父母的督促下练毛笔字、背诵古文和诗词，接受中国传统的人文教育；十一二岁就开始阅读《三国演义》《西游记》《水浒传》《镜花缘》等古典小说，还翻阅了二十余本一套的《历朝通俗演义》（自两汉至民国元年），在修习中国传统文化的同时养成了读书自学的好习惯。同时，他主动地接受新文化的洗礼，阅读鲁迅的《狂人日记》、巴金的《新生》《灭亡》《家》《春》《秋》、苏联革命年代的小说《表》《面包》《士敏土》、艾思奇的《大众哲学》、斯诺的《西行漫记》以及德国柯勒惠支的版画等，努力地接受进步思想的启蒙和教育。他早年曾在南京钟英中学、上海正始中学、上海大同大学附中读书，期间满怀爱国热情地投入上海的抗日救亡运动，不仅参加了示威游行，还跟同学一起参加"赴京请愿团"到南京请愿，被军警在无锡拦截、押送回上海。可见，先生年轻时并不只是一个埋头读书的书生，而且是一个热爱国家、关心政治、并且勇于投身社会的热血少年。

1939年，朱德熙先生考取昆明西南联合大学物理系，比杨振宁先生低一班。后来，由于受到清华大学哲学系研究生朱南跣和徐孝通的影响和启发，于第二年（1940年9月）转入中文系学习。在中文系学习期间，他受到了罗常培、唐兰、陈梦家等教授的教导和赏识，学问进步很快。期间休学过两年，延至1945年毕业。毕业后，先生曾在昆明中法大学中文系任教，并加入了中国民主同盟。1946年应清

---

① 这一部分主要参考《朱德熙先生纪念文集》（语文出版社，1993年）中的《朱德熙先生生平》、李荣《朱德熙》、朱德熊等《忆大哥》等文章。

华大学中文系主任闻一多先生的聘请，去清华大学中文系任教。1952年，因院系调整先生调入北京大学中文系工作，并应邀赴保加利亚索菲亚大学任教。1955年回国，此后他一直在北京大学中文系工作1979年晋升为教授。朱德熙先生先后担任过北京大学中文系副主任，北京大学计算语言学研究所所长，北京大学副校长兼研究生院院长，中国语言学会副会长、会长，世界汉语教学学会会长兼《世界汉语教学》主编，中国古文字研究会理事，国务院学位委员会委员，国家语言文字工作委员会委员，国务院古籍整理规划小组顾问，中国大百科全书总编辑委员会委员，第五、六届中国民主同盟中央委员会委员，第六、七届全国人民代表大会代表，第七届全国人民代表大会常务委员会委员、文教委员会委员等职。在这繁重的社会活动之外，他仍孜孜不倦地从事学术研究和教学工作，并不断有令人耳目一新的成果问世。

朱德熙先生以其精湛的汉语语法和古文字方面的研究成果而蜚声于国内外的汉语语言学界，除了保加利亚之外，还先后赴美国、法国、泰国、香港、新加坡、澳大利亚等国家和地区讲学、合作研究或出席国际会议。1986年，法国巴黎第七大学授予他荣誉博士学位。

朱德熙先生为人谦虚方正、耿直又不失厚道，他思想开阔、兴趣广泛：一方面学习国外先进的语言学理论和方法，另一方面熟读中国古书、研究古文字；一方面潜心学术研究，另一方面唱昆曲、吹笛子，既会研究又会娱乐。先生热爱教育事业，不断地奖掖后进，培养了一大批优秀的汉语语言学研究和教学人才。

1991年12月，朱德熙先生病重，被确诊为不治之症。1992年7月19日清晨6时6分，先生在美国斯坦福大学医院逝世，享年72岁。

朱德熙先生的一生坎坷多艰，早年外敌入侵和内战祸乱使他颠沛流离、生活不宁，壮年政治运动接连不断使他没有平静的书斋，晚年病痛的折磨使他难以把最后一篇论文写完。但是，先生在艰难中发奋读书、矢志学问、坚韧不拔，把生命融入学术，取得了令世人瞩目的学术成就。

2 四十余载治语法 为"的"消得人憔悴

1956年,朱德熙先生发表《现代汉语形容词研究》(《语言研究》第1期),全面而系统地用分布分析的方法来说明:形容词的简单形式和复杂形式在语法功能上有一系列的区别。所谓简单形式,指的是形容词的基本形式,包括单音节形容词和一般的双音节形容词。例如:

大、红、多、快、好、干净、大方、糊涂、规矩、伟大

所谓复杂形式,指的是形容词的各种重叠形式、带后缀的形容词、偏正式的夸饰类的形容词、以形容词为中心的词组。例如:

小小儿、远远儿、老老实实、干干净净、糊里糊涂、古里古怪、黑乎乎、慢腾腾、脏里瓜唧、白不雌列、雪白、通红、挺好、又高又大

为了方便,文章称形容词的简单形式为甲类成分,称形容词的复杂形式为乙类成分。他指出,从意念上看,甲类成分表示的是单纯的属性,乙类成分表示的是这种属性的状况或情态、它跟说话人对于这种属性的主观估价作用发生联系(即包含着说话人的感情色彩在内)。更引人入胜的是,这种意念上的区别完整地反映在甲、乙两类成分的语法功能上——不论在什么样的环境里,这两类成分始终表现着互相对立的倾向。文章分别从定语、状语、谓语、补语四种位置上来观察甲、乙两类成分在语法功能上的区别:

第一,甲类成分充当的定语是限制性的,如在"白纸"里,我们用"白"这种属性来限制"纸"这个类名,得到一个新的类名"白纸";乙类成分充当的定语是描写性的,如在"雪白的纸"里,"雪白的"不是用来作为分类的根据,而是用来描写"纸"的状况或情态的。甲类成分充当的定语跟其中心语是互相选择的,二者不能任意替换;乙类成分充当的定语跟其中心语的选择关系相对自由,只要二者的词汇意义不抵触就行。例如:

凉水 ～ *凉脸 ～ 冰凉的脸

薄纸 ～ *薄灰尘 ～ 薄薄的灰尘

第二,由形容词充当的状语表示的是动作的方式或状态,属于描写性的、而不是限制性的。因此,甲类成分一般不宜于作状语,而乙类成分(特别是其中的重叠式)则经常担任状语这种职务。

第三,甲类成分作谓语的句子,含有比较或对照的意思,因此往往是两件事对比着说的;乙类成分作谓语的句子,没有比较或对照的意思,因此可以独立出现。例如:

今儿冷,昨儿暖和。～ 今儿怪冷的。

第四,甲类成分作补语的句子,含有比较或对照的意思,因此往往是两件事对比着说的;乙类成分作补语的句子,没有比较或对照的意思,因此可以独立出现。例如:

站得高,看得远。～ 他的嘴张得大大的。

这种形式跟意义互相渗透、互相验证的研究方法,是对美国结构主义只重形式、不顾意义的重大改进。被赵元任(1968)《中国话的文法》誉为:"到目前为止,讨论中国形容词的文章,最好的还算这一篇"(见丁邦新译本第566页,河北教育出版社,1996年)。就在研究甲、乙两类形容词的语法功能的差别时,朱先生敏锐地发现:"甲的"(甲类成分加"的")跟"乙的"(乙类成分加"的")的语法性质不一样,前者是形容词性的、可以受副词的修饰,后者是体词性的、可以受数量词或指示词修饰。例如:

脸上永远红扑扑的～ * 脸上永远红的

* 一个大大的～ 一个大的

进一步,朱先生发现"甲的"跟"乙的"中的"的"虽然在形式上没有区别,但是它们的语法性质很不一样:前者有体词化的作用,后者没有这种作用。

那么,北京话里读"的"的形式到底代表几个语素呢?对此,朱德熙先生采用严格的分布分析方法进行深入的研究,并在1961年发表《说"的"》(《中国语文》12月号)中,公布了这种令人难以置信、却又不得不相信的结论:通过比较不带"的"的语法单位——假定为x——跟加上"的"之后的格式"x 的"在语法功能上的差别,由此分离出"的"的性质来;即根据不同的 x 加上"的"之后形成的格式("x_1 的"、"x_2 的"等)在功能上的区别,把"的"字分析为三个不同的语素:

“的$_1$”是副词性语法单位的后附成分，“副词＋的$_1$”仍是副词性的，即只能作状语。例如：

天渐渐(的)黑了 忽然(的)门被风吹开了

“的$_2$”是形容词性语法单位的后附成分，可以通过单音节形容词的重叠形式AA(儿)(记作R)来证明。R分两类，一类一定要后附“的”，这种R记作R_a。“R_a的”是形容词性的，可以单说、作谓语、补语、定语和状语。例如：

红红儿的 脸红红的 抹得红红的 红红的脸 热热的喝下去

另一类R只能作状语，是副词性的，记作R_b。但是，“R_b的”是形容词性的。例如：

好好拿着 ～ 什么都好好的、说得好好的、好好的东西、好好的拿着

可见，“的$_2$”有把形容词的重叠形式转变为形容词性的语法单位的功能。“的$_3$”是名词性语法单位的后附成分，形容词、动词、名词加上“的$_3$”后在语法功能上是名词性的，可以作主语、宾语、定语和体词性谓语。例如：

白的好 不要白的 白的纸 这张纸白的

这种把带“的”的格式在语法功能上的异或同归结为后附成分“的”的异或同的研究方法，引起了当时语言学界极大的关注和极为激烈的争论。为此，朱先生在1966年发表《关于〈说“的”〉》(《中国语文》第1期)，一方面澄清各种误解，另一方面进一步阐明这种方法的实质和根据。限于当时的政治环境，先生矢口否认这是描写语言学派的方法，声称这是传统语言学对付印欧语系各种语言时沿用的老办法。值得重视的是，在这篇文章中，朱先生联系历史，把唐宋时期带“底、地”的格式分为三类：

(1)“x底”，它只能作主语、宾语、表语、定语，不能作状语，是名词性成分；

(2)“x地$_2$”，它能作谓语、状语、定语，是形容词性成分；

(3)“x地$_1$”，它只能作状语，如“陌地、平白地”，是副词性成分。

据此，他把“底、地”区分为三个语素：“地$_1$”是副词的后附成分、“地$_2$”

是形容词的后附成分、“底”名词性单位的后附成分，并认为：现代汉语的“的$_1$、的$_2$、的$_3$”是分别从唐宋时期的“地$_1$、地$_2$、底”演变来的，历史事实支持他的这种分析。

此后，由于“文化大革命”的政治干扰，朱德熙先生被迫中断语法研究；直至1976年以后，他才有机会重新从事语法研究。

1978年，中国学术界刚从十年浩劫中苏醒，朱德熙先生就在刚复刊的《中国语文》(第1、2期)上发表《“的”字结构和判断句》，引进动词的“向”、“潜主语”、“潜宾语”等概念来讨论“的”字结构的语义所指和歧义指数、分析由“的”字结构组成的五种判断句。1980年，朱先生发表《北京话、广州话、文水话和福州话里的“的”字》(《方言》第3期)，指出广州话的三个“的”读音不同(分别写作：咁、哋、嘅)，语法功能不同，显然是三个不同的语素；北京话的三个“的”同音，分析起来要困难得多；但是，广州话、文水话和福州话里“的$_1$、的$_2$、的$_3$”三分的局面以及历史上“地$_1$、地$_2$、底”的区分都支持他对北京话的“的”所作的分析，尽管这四种方言里相对应的“的$_1$、的$_2$、的$_3$”的来历不一定都相同。也就是说，经过十几年的摸索和思考，先生终于找到了一条贯通方言和历史的现代汉语语法研究的路子。1983年，朱先生发表《自指和转指——汉语名词化标记“的、者、所、之”的语法功能和语义功能》(《方言》第3期)，引进句法成分的“提取”、“缺位”和名词化形式的“自指”、“转指”等概念，分析了现代汉语的“的”和古汉语的“者、所、之”等名词化标记的性质，并且从语法功能和语义功能两方面比较了它们的异同。1991年，朱先生写成《“的”字的方言比较研究》，[①] 利用汉语18个方言点的材料，讨论状态词的名词化、“的$_1$”和“的$_2$、的$_3$”的关系、状态词的名词化形式的指称功能和陈述功能、有没有专作定语标记的“的$_4$”、文章还利用《祖堂集》等新语料对“地、底”的分布及其跟“的$_1$、的$_2$、的$_3$”的源流关系进行了考察。

1992年，在生命的最后岁月，朱德熙先生强忍着病痛，在《“的”

① 该文曾以《汉语助词“的”的跨方言比较研究》为题，提交在康乃尔大学举行的第三届北美洲汉语语言学会议(1991年5月3日—5日)。关于这次会议，请看《国外语言学》1991年第4期的会议报道(第44—45页)。

字的方言比较研究》这篇文章的基础上撰写《从方言和历史看状态形容词的名词化兼论汉语同位性偏正结构》(未完成稿，发表在《方言》1993 年第 2 期)。文章考察了分属六个大方言区的十种方言里的状态形容词的后缀(即的$_2$)的语音形式、语法分布及其名词化时与名词化标记(即的$_3$)的组合关系。文章着重指出:(1) 在那十种方言里，状态形容词充任定语时必须通过加“的$_3$”的办法名词化;由于“的$_3$”除了有名词化的功能外，还有语义上的转指功能，这样造成的偏正结构都是同位性的。(2) 同位性偏正结构在现代汉语各类名词性偏正结构里所占的比重极大，这种局面开始形成于唐宋之际“者”字嬗变为“底”字的时期;由“底”字组成的新的同位性偏正结构的兴起是汉语语法史上的一件大事。尽管先生没来得及写出对同位性偏正结构的全部见解，但是他还是先写出了余论，对自己三十多年研究“的”字的经历和得失作了一番发人深省的总结:1961 年写《说“的”》时没想到跟广州话等方言作比较，否则很容易得到“的”字应该三分的结论;当时批评《说“的”》的文章也只是说它不提历史、不说它不提方言，因为那时很多人心目中都没有方言语法比较这回事;1980 年写《北京话、广州话、文水话和福州话里的“的”字》，主要是想说明方言事实也支持《说“的”》的分析，却发现了方言里状态形容词修饰名词的时候要名词化的事实，但没想到应该回头去考察历史，看看这个现象在文献里是否有反映;直到 1989 年重新拣起这项工作时才去查考历史，结果发现历史事实跟方言情况完全一致——状态形容词作定语的时候也必须名词化。经过这三十年的循环，朱先生对方言语法研究、历史语法研究和标准语语法研究的密切关系最终有了深切的体会，并用这篇生命的压轴之作向学术界展示这种远见卓识。

纵观朱德熙先生一生中一些主要的研究课题，我们可以发现它们大多跟“的”字相关;一个小小的“的”字，牵动着汉语语法的全局，耗尽了先生毕生的心血。[①]

---

① 朱德熙先生说:“我写一千字，起码要用掉两三千字的稿子，一篇文章写完，就像是得过一场病似的”(见《朱德熙先生纪念文集》中林焘先生的《哭德熙兄》，第 86 页)。这大概可以作为我们这儿几句话以及本节标题的脚注。

3 筚路蓝缕辟蹊径 融汇中西铸新篇[①]

50年代初,朱德熙先生跟吕叔湘先生合写《语法修辞讲话》(1951年6月5日起在《人民日报》连载)。从此,朱先生把很大的时间和精力投放到语法研究上。在进行语法研究的实践中,朱先生努力从汉语语法的事实出发,吸收国外结构主义语言学的新理论、采用美国描写语言学的新方法,不断地探索汉语语法研究的新途径、开辟汉语语法研究的新领域、并逐步创立汉语语法学的新体系。先生在1956年的《现代汉语形容词研究》中,已经全面而系统地用分布分析的方法来证明形容词的简单形式和复杂形式在语法功能上有一系列的区别,为对汉语语法现象进行分布分析作出了示范。在1961年的《说“的”》中,先生更是把分布分析的效用推到了极致,把一个普通常用的“的”区分为功能迥异的三个语素;这种结论一方面使人难于接受,一方面又使人不得不信服,着实是令人耳目一新。更有意思的是,在1966年发表《关于〈说“的”〉》中,先生在说明研究方法时,把结构语法学的分布分析方法跟传统语法的渊源关系交代得清清楚楚:英语合成词x-ly有副词(如:partly, roughly, determinedly)、形容词(如:cowardly, lowly)两种词类,其中的x有名词(如:part, coward)、形容词(如:rough, low)、动词(如:determined)三种词类;如果根据x的功能来区分-ly的话,可以分成三个不同的-ly来:一个只能在名词后头出现,一个只能在形容词后头出现,一个只能在动词后头出现;但是,这种分析方法不能完全反映-ly的语法作用。如果根据x-ly的功能来区分-ly的话,可以分出两个不同的-ly来:一个造成形容词,一个造成副词;这种分析方法能够反映-ly的语法作用,即形容词化和副词化。传统语言学采用的是后一种方法,这种方法的实质是把带-ly的格式的功能上的异或同归结为-ly的功能的异或同。而《说“的”》的根本方法就是把带“的”的格式在功能上的异或同归结为

① 这一部分主要参考《朱德熙先生纪念文集》中陆俭明先生的《朱德熙先生在汉语语法研究上的贡献》等文章。

"的"的功能的异或同。从中,我们固然可以看出先生的雄辩,更可以领略他对于新旧两种分析方法的内在联系的洞察力。的确,分布分析这种描写语言学的新方法是建基于传统的语法分析方法之上的,是对传统分析方法的理论化、程序化。事实上,分布分析的理论前提是语言构造的层次性;因此,只有先对语言形式进行层次分析才能进一步对其中的某些语法形式进行分布分析。朱先生在《现代汉语形容词研究》中就碰到"白的纸"应该二分(白的/纸)还是三分(白/的/纸)、"歪戴着、白跑了"该怎样切分(是"歪/戴着、白/跑了"还是"歪戴/着、白跑/了")等问题,在《说"的"》中又碰到"S的M"应该二分(S的/M)还是三分(S/的/M)、"真的、善的、美的东西"该怎样切分等问题。但是,当时的汉语语法研究基本上是在传统语法的框架中进行的,分析句子用的是主谓宾定状补六大成分一字排开的成分分析法。为了扭转这种局面,先生在1962年的《论句法结构》(《中国语文》8—9月号)中,详细地论述了语法构造的层次性,介绍了层次分析的基本的操作程序,说明了层次分析对于分化歧义结构的作用,给当时的汉语语法学界劈头猛浇了一场及时雨。

在1961年的《说"的"》中,朱先生就开始尝试变换分析法。先生用这种动态的、能把有关句式系联起来的方法,来证明"我会写的"这类结构中的"的"是名词性语法单位的后附成分"的$_3$",而不是一般所谓的语气词。在1962年《论句法结构》中,朱先生专辟一节来介绍变换分析,并以汉语实例说明只有通过变换关系才能找出严格意义的同构格式来——因为狭义同构的语法形式内部并不是完全一致的,这种不一致性可以从它们对于特定的变换式的不同反映上看出来。例如:

台上坐着主席团 → 主席团坐在(得)台上

台上唱着戏 → *戏唱在(得)台上

可见,符合狭义同构条件的格式"处所词+动词+着+名词"并不是真正的同构。在以后发表的文章中,朱先生更为广泛和娴熟地运用变换分析方法来分析一些利用成分分析法和层次分析法所无法处理或难以解释的语法现象,特别是歧义现象。在1986年的《变换分析的平行性原则》(《中国语文》第2期)中,先生系统地总结了变换分析

法的理论原则、应用步骤，详细地阐明了变换式矩阵里的句子之间的四种平行关系。不仅为我们更加深入地分析汉语语法、巧妙地揭示隐蔽的语法规律提供了新的方法，而且为我们作出了怎样使用变换分析法的指导性的说明。

在1978年的《“在黑板上写字”及相关句式》（《语言教学与研究》试刊，第三集）中，朱先生用变换分析法来分化一类同形异义句式。例如：

黑板上写着字→（把）字写在黑板上

屋里开着会→*（把）会开在屋里

并且用语义特征分析法来说明造成这种句式同形异义的原因——动词“写”等有〔十状态〕〔十附着〕的语义特征，动词“开（会）”没有这种语义特征。后来，先生又多次修改这篇文章，不断地完善语义特征分析法；第一次修改稿发表在1981年的《语言教学与研究》（第1期），第二次修改稿收入论文集《语法丛稿》（上海教育出版社，1990年）。在1979年的《与动词“给”相关的句法问题》（《方言》第2期）中，朱先生继续用变换分析法分化跟动词“给”相关的句式，并用主要动词是否包含〔十给予〕或〔十取得〕等语义特征来作出解释。如果说，变换分析法为分化同形句式提供了可操作的形式程序；那么可以说，语义特征分析法为解释同形句式为何异义提供了直观的意义根据。变换分析和语义特征分析的配套使用，使汉语语法研究走上了形式和意义互相结合、互相渗透、互相验证的道路。

朱德熙先生不仅在语法研究的具体方法上不断创新，而且在研究语言的观察角度和宏观思路上也不断反思、锐意革新。他在1985年为桥本万太郎的《语言地理类型学》中译本所写的序中，一方面肯定了德·索绪尔区分共时的和历时的语言研究方法的学说，给二十世纪的语言研究带来的深刻的积极影响；另一方面也明确指出这种学说的消极影响：把对语言的历史研究和断代描写截然分开，看成是毫不相干的东西。朱先生批评了这种思潮对汉语研究的消极影响：研究现代汉语的人往往只研究普通话，不但不关心历史，而且把方言研究也看成隔行。为了彻底地改变这种风气，先生身体力行，将共时的各种方言（包括北京话和普通话）之间的比较研究跟历时的古今汉

语语法之间的比较研究结合起来，先后对“的、者、所、之”等虚词、重叠式象声词、反复问句等语法问题进行了跨越方言和贯通古今的对比研究，写出了令人眼界大开的《北京话、广州话、文水话和福州话里的“的”字》、《自指和转指——汉语名词化标记“的、者、所、之”的语法功能和语义功能》、《“的”字的方言比较研究》、《从方言和历史看状态形容词的名词化兼论汉语同位性偏正结构》、《朝阳话和北京话重叠式象声词的构造》（《方言》1982年第2期）、《汉语方言里的两种反复问句》、《“V-neg-VO”与“VO-neg-V”两种反复问句在汉语方言里的分布》。这一系列开创性的研究工作，为汉语语法研究开辟了一条贯通共时和历时的路子，扩大了我们的视野、拓宽了我们的思路，并使汉语语法研究走上了全方位、多视角的道路。

在长期的语法研究实践中，朱德熙先生一方面仔细观察各种语言事实、深入挖掘各种语言现象背后隐藏的规律；另一方面努力学习各种先进的理论、大胆尝试各种新颖的方法，在汉语词类划分、汉语语法的特点、汉语语法学体系、语法成分之间的结构关系、语法结构之间的关系和转换过程等方面提出了一系列富有创见又启人深思的理论观点。众所周知，由于汉语没有形态变化，因而汉语到底有没有词类分别、如果有那么应该根据什么标准来划分等一直是困扰语法学界的难题。早在50年代，朱先生已经明确地指出：“我认为划分词类的基本根据应该是词的语法功能。……在形态丰富的语言里可以根据形态划分词类，……形态不过是功能的标志而已”。[①] 在八十年代出版的《语法讲义》和《语法答问》中，先生坚持并发展了这种观点，指出：“汉语不像印欧语那样有丰富的形态。因此给汉语的词分类不能根据形态，只能根据词的语法功能。……一个词的语法功能指的是这个词在句法结构里所能占据的语法位置”（《语法讲义》§2.1）。这在理论上为汉语词类划分工作提供了指导性的原则。在《语法答问》中，朱先生对比了汉语和英语等印欧语言的主要差别，指出由于

① 详见《北京大学1959年五四科学讨论会讨论汉语实词分类问题的报告和发言》中朱德熙先生的发言，《语言学论丛》第四辑，上海教育出版社，1960年。又《朱德论文集》第2卷，商务印书馆，1999年。

汉语缺乏形态变化，造成汉语语法有两大特点：(i)汉语词类跟句法成分之间不存在简单的一一对应的关系，而是呈现出一对多、多对一、多对多等错综复杂的对应关系，进而指出词类转化、词无定类学说的错误根源在于没能认识汉语词类的多功能性；(ii)汉语句子的构造原则跟词组的构造原则基本上是一致的，不像英语那样句子和子句是一套构造原则、词组是另一套构造原则，因此可以在词组的基础上来描写汉语句法、建立一种以词组为基点（本位）的语法体系。在 1980 年发表的《汉语句法里的歧义现象》(《中国语文》第 2 期)中，先生在讨论多义句式的分化依据时，明确地指出句法成分之间有两种不同的语法关系：一种是显性的语法关系，比如主谓、述宾、偏正等结构关系；一种是隐性的语法关系，比如动作和施事、受事、工具等语义关系。区分这两种语法关系，为解决汉语主语、宾语的区分和界定提供了理论指导。在句法结构的类型上，先生在《语法讲义》中开创性地区分出粘合式和组合式两种结构类型：粘合式结构的组成成分都是单个的词，整个结构在功能上相当于一个词；组合式结构的组成成分一般不是单个的词，其中的结构成分之间关系比较松散。建立了粘合、组合的概念，可以更好地概括一些语法现象、解释一些语法规律，比如多项定语的顺序、哪些述宾结构可以直接作名词的定语等问题。在 1982 年的《语法讲义》中，先生提出了指称和陈述两个新概念。指称就是名词性成分在意念指谓事物，指称形式可以用“什么”来指代；陈述就是动词性成分在意念上指谓事件或状态等，陈述形式可以用“怎么样”来指代。在《自指和转指》一文中，先生系统地用这两个概念来说明不同的句法结构之间的转化关系：“的”加在动词性成分（记作 VP）之后，原来表示陈述的 VP 就转化成表示指称的“VP 的”了。比如：

〔小王〕开车 → 开车的(＝小王)

小孩画〔画儿〕→ 小孩画的(＝画儿)

在讨论陈述形式向指称形式转换（即名词化）的时候，先生十分娴熟地运用了国外语言学理论中的句法缺位和成分提取等概念，用以说明名词化的指称形式在语义上有自指和转指两种情况。其中，自指是名词化造成的指称形式跟原来的动词性成分所指相同，比如“(小

王)开车的技术"中"(小王)开车的"跟"(小王)开车"所指一样;转指是名词化造成的指称形式跟原来的动词性成分所指不同,比如"〔〕开车的(人)"中"〔〕开车的"跟"〔〕开车"所指不同。先生十分敏锐地指出,自指的名词化形式中可以有、也可以没有缺位,转指的名词化形式中一定没有缺位。这为我们进一步研究汉语句法结构之间的转换关系、研究"的、者、所、之"等虚词的句法功能和语义功能提供了合用的理论概念和分析方法。值得一提的是,在1978年的《"的"字结构和判断句》中,先生首次明确地提出汉语动词"向"的观念,并用以解释"的"字结构的歧义指数;这一方面为我们研究歧义现象、描写句子成分之间的支配和从属关系提供了新的角度和工具,另一方面直接推动了汉语配价语法的蓬勃展开。

因此,我们十分赞同陆俭明先生的评论:朱德熙先生是我国思想最活跃、最富有创新精神的语法学家之一,是汉语语法研究的带头人和领路人。[①]

4 一生钟情古文字 探幽发微结硕果[②]

朱德熙先生在大学期间就对古文字产生了浓厚的兴趣,课余时间专心攻读《说文解字》,用的本子是扫叶山房石印的段玉裁《〈说文解字〉注》,经常跟李荣先生等同学讨论说文。他还听了唐兰先生讲的《说文解字》和《古文字学》两门课,毕业论文的选题就是关于甲骨文研究方面的,并被闻一多教授亲自批为甲等。自1947年到1948年,朱先生在北平《新生报》的语言与文学版上连续发表《读古文字小记》(2篇)和《楚器研究》(3篇)等考释古文字的文章。其中,关于战国楚器铭文的几篇,后来改写成《寿县出土楚器铭文研究》发表在郭沫若先生主编的《历史研究》创刊号(1954年第1期)。在这篇文章中,朱先生通过细密的论证,考释出写法奇诡的"佳"字和当时许多人

① 详见陆俭明《朱德熙先生在汉语语法研究上的贡献》,见《朱德熙先生纪念文集》。

② 这一部分主要参考《朱德熙先生纪念文集》中李学勤先生的《朱德熙先生战国文字研究的贡献》、裘锡圭先生的《朱德熙先生在古文字学方面的贡献》等文章。

不能辨识的从“佳”的“集”字，并很快就得到了普遍承认。另外，文章把楚器铭文中的“王句”和古印中的“夫句”释读为“王后”和“太后”，这些都是很好的见解。这些文章的水平都明显超过以往对寿县李三孤堆铭文的考释，当时便引起了学术界的广泛注意。1858年，朱先生又在北京大学中文系编辑的《语言学论丛》第二辑上发表《战国记容器刻辞考释四篇》，提出了不少很好的意见。正是这两篇文章奠定了朱先生在战国文字研究领域中的突出地位。

到了60年代，尽管当时的政治形势不利于学术研究，但是朱先生由于对古文字有发自内心的爱好，并没有中止这方面的研究。在70年代，先生参加了马王堆1号汉墓遗册、银雀山汉墓竹书、马王堆3号汉墓帛书、望山楚墓竹简和平山中山王墓铜器铭文的整理研究工作，作出了很大的贡献。同时，朱先生在《文物》《考古学报》《古文字研究》《方言》上发表了多篇考释战国文字的论文，涉及的资料包括楚简、楚帛书、玺印、陶文、盟书和铜器铭文等方面，提出了一些极为精辟的考释结论，并为学者们所普遍接受。

战国文字是很难研究的，因为战国时代“言语异声，文字异形”（说文·叙），又是文字剧烈变化的一个时代。秦灭六国以后，“罢其不与秦文合者”（说文·叙），使得汉代的人已难以辨识。并且，战国文字里的很多字跟各种古文字里相应的字，在字形上几乎完全失去了联系，令许多学者望而却步。历代学者解释玺印、货币等战国文字，有好多误解；而且这些误释流传甚广，几乎成了公认的说法。朱先生的不少研究就是针对传统误释的，他在占有大量材料的基础上作出周密的论证，其结果往往出人意表但又令人信服。因为先生善于深入细致地分析字形，精确地揭示该字的形体演变的复杂过程，使人看清这一字形就是该字的异体。

朱先生研究古文字还有一个特点，就是善于把语言学的方法运用到古文字学的研究上来。先生坚持语言学的观点，非常注意文例；不但要求自己的考释在字形上站得住脚，而且要求在语法、语义上也站得住脚。他时常根据一个古文字在语句中的语法地位，来判断它有可能是哪个字或不可能是哪个字。由于朱先生具有上述优点，加上他的文章语言简练、条理清楚、分析透彻，因而他的考释文章给人

耳目一新的感觉。正是由于他和裘锡圭先生在70年代以后的合作，共同促进了战国文字的研究，并使战国文字研究成为中国古文字学的一个独立的分支。

朱德熙先生毕生热爱古文字，那是一种发自内心的、情不自禁的挚爱。有两件小事或许可以为证：1975年，先生在文物出版社整理湖南长沙马王堆出土的竹简；这时，恰逢先生的外孙女降生，他就兴致勃勃地给外孙女起了个单名“简”；[①]在先生晚年，有人问他：“在语法和古文字两头，您哪方面成就更大一点？”先生沉吟片刻，笑着说：“大概差不多吧。”含糊的回答，实在是一种真情的流露。

5 哲人已骑黄鹤去 薪火熊熊有传人

朱德熙先生离开了我们，带着他的遗憾、带着他的未竟之作、带着他那远大的学术抱负、带着他那过人的才智和广博的学识；但是，朱先生的学术思想留给了我们、朱先生的创新精神留给了我们。令人欣慰的是，直接或间接地受过朱先生教诲和学术影响的学者，他们散布在祖国各地、乃至世界各地；他们正在以不同的方式发扬先生开创的学术思想，努力把汉语语言学的研究推进到一个更高的水平。在朱先生生前工作的北京大学中文系，在现代汉语语法、近代汉语语法、方言语法、古文字等先生热爱的研究领域，均有实力比较雄厚的研究队伍，并形成老中青三代井然有序的学术梯队，先生开创的学术事业可以说是后继有人、并必将更加蓬勃地发展。

（原载袁毓林编《朱德熙选集》，东北师范大学出版社，2001年）

① 详见《朱德熙先生纪念文集》中朱襄的《我们想念你，爸爸》，第67页。

后记

收入本书的18篇文章，都是我在90年代中期以后陆续写成的。其中，大部分是作为教育部"十五科研规划第一批（博士点基金）项目"——"面向信息抽取的语义标注研究"的子课题，而陆续完成的（项目批准号：01JB740006）。现在，我把它们收集在一起，根据文章的内容，大致分成4编，以便读者阅读。在这里，我要感谢教育部给我提供这笔基金，使我能够在一个比较优越的环境中进行研究、从容地写作。我还要感谢北京大学社会科学研究部提供了相应的配套经费，使我的研究有了比较充分的物质保障。虽然加在一起的资金并不多，但是足以让我感受到我们这一代学者的幸运，能够享受改革开放带来的伟大成果。这也促使我知恩图报，努力在自己所从事的专业领域中作出成绩，尽可能让自己的研究多少带有一点技术色彩，希望为国家的经济建设献出一点绵薄之力。

大家知道，计算语言学的研究有不同的思路（approach）；而这跟不同的研究者对这门学科的理解，特别是他们的研究取向（orientation）和知识背景有关。我比较喜欢那种对于语言研究和自然语言的计算机处理两头都有启发性的路子，于是免不了要走认知主义的道路，并坚持认知的本质是计算的观念；在此基础上，逐步形成基于认知并面向计算的语言研究的路子。当然，我不反对其他路子的计算语言学研究。这反过来也可以解释，为什么我对语言的认知研究跟一般的认知语言学或认知语法会如此大异其趣。

在学习和研究计算语言学的过程中，我先后得到黄昌宁、罗振声和董振东等老师的帮助和鼓励；跟白硕、王培、金茂兵、孙茂松、周明、姬东鸿和周强等学友的讨论，也使我大开眼界。特别是陆俭明老师鼓励我于1998年给研究生开设《计算语言学》课程，促使我对计算语言学的各种研究路子和教材体系作了系统的梳理。在此，谨向他们表示诚挚的谢意。

在这些文章的写作和修改过程中，先后得到陆俭明老师和顾阳、郭锐、沈培、詹卫东、徐刚等学友的帮助和指正，还得到方梅、刘丹青、张旺熹、靳光瑾、叶青和申坚等先生的指正；其中《走向多层面互动的汉语研究》一文是在日本讲学期间写成的，承蒙东京大学的柯理思教授、御茶之水女子大学的相原茂教授和森中野枝助手提供诸多资料上的帮助。在此，谨向他们表示诚挚的谢意。在这里，我特别要感谢《中文信息学报》编辑部的曹右琦等老师和多位匿名评审老师；正是他们的鼓励和帮助，使我的三篇文章能够在《中文信息学报》上发表，可以直接向中文信息处理界的广大人士请教。

最后，我要感谢陆俭明老师对我的关心和教诲；感谢他在百忙之中拨冗作序，鼓励有加。感谢冯志伟老师对我的关心和鼓励，并欣然答应作序。感谢同事詹卫东先生在该课题的申报、文章中逻辑公式的推敲、一直到图表处理等诸多方面提供的大量帮助。我还要感谢北京大学出版社的热情支持。

由于收入本书的各篇文章是在不同时期和不同的地点写成的，因而所引文献的版本、行文的格式体例、所用的术语乃至观点可能前后不一，现在也难以全部统一；谨此，向广大读者致歉。书中的谬误和纰漏，敬请各位读者和行家不吝指正。

2007 年金秋于北京蓝旗营